University of Tartu

School of Economics and Business

Sander Sõna

PREDICTING INNOVATING COMPANIES IN ESTONIA BY ANALYSING MANUFACTURE COMPANIES WEBSITE DATA

Master thesis

Supervisors: Jaan Masso, Rajesh Sharma, Priit Vahter

2020

Name and signature of supervisor Jaan Masso, Rajesh Sharma, Priit Vahter

Allowed for defence on..... (date)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced. …….. (signature of author)

TABEL OF CONTENTS

ABSTRACT

This paper is testing a new approach for understanding novelty in companies through webpages. Specifically, the center of the study is the Estonian manufacturing sector. I used traditional firm-level innovation indicators from the questionnaire-based Community Innovation Survey (CIS) to train the neural network classification model with labeled (innovative and non-innovative) web texts. This approach shows results and it offers highly cost-efficient alternatives to understand innovation in addition to the existing innovation indicators. This method allows learning new innovation indicators just from companies' website texts.

Keywords: Manufacturing, innovation, Community Innovation Survey, machine learning, neural network.

Common European Research Classification Scheme (CERCS):
S180 - Economics, econometrics, economic theory, economic systems, economic policy,
S191 - Market study,
P160 - Statistics, operation research, programming, actuarial mathematics.

INTRODUCTION

Efforts have been made to measure innovation since after the second World War when Joseph Schumpeter first brought out the concept. This was a time when innovation statistics as R&D (research and development) started to be collected. The main push was given in 1966 when Chris Freeman founded the University of Sussex in the United Kingdom. Under his leadership, the university focused on the innovation studies that helped, to get a solid ground in this area. Of course, the term "innovation "was not widespread at that time. Innovation topic was mostly related to science studies or science policy studies (Clausen *et al.* 2012: 1254-1255). So, it was natural that studies were conducted on data, like R&D or patents, what was collected in previous decades. Although there have been studies until the last decade that tried to measure innovation with company accounting indicators, there is more evidence that qualitative information (like surveys and texts) cannot be ignored. (Smith 2009: 151-152; Klette 1996: 509-510; Henderson and Cockburn 1994: 12-13; Argyres and Silverman 2004: 938-939; Clem *et al.* 2004: 409, 418-420; Lhuillery *et al.* 2016: 7) Survey-based measures, like Community Innovation Survey (CIS), that investigate the perceptions of managers about innovation are already in common because it is carried out in European Union member states every two years. (Community Innovation..., 2020). The reasoning for this is that quantitative data may limit innovation definition and understanding of it. To fully understand this topic there is also a need to combine and test different types of researches, like e-administrative records, internet data, or social media. (Sauermann and Roarch 2013: 276-279; Geuna *et al.* 2015: 1647, 1650-1652) The difficulty level of analyzing qualitative data has been decreased over the years because of the increase in computer computing power (Schmidhuber 2015: 86-87; Kinne and Lenz 2019: 1). So, innovation studies can and should use new methods to discover its phenomenon.

Companies are trying to express themselves in different ways to clients and partners to spread their information. Nowadays one of the main company's communication channels is the internet and most importantly their homepage. The homepage is like an

organization's business card, so it must be well informed. At the same time homepage is the first place where clients and partners go to get more information. So, webpages are important for both sides (companies themselves and for external stakeholders). That is more why the website data can be used for the paper to predict the company's innovation (Fisher *et. al.* 2007: 253; Gök *et.al.* 2015: 654). There is a need to research new innovation indicators. Traditional innovation indicators may have problems with coverage, time, and money (Kinne and Lenz 2019: 1; OECD, 2015: 26-28; Kinne and Axenberg 2018; 3-4). Hence, most of the data on the internet is free and if there is a possibility to get data, then it is used. But the main point with coverage goes well with time too. These kinds of studies do not need interactions with other participants or contacting the people. It all depends only on where and how to mine data on the web. And because it is mostly free (except a need for database-specific classification data) it does not need interviewers or other types of data collecting effort. There have been successful tries to let the machine algorithm to learn about the company's homepages. Germans Kinne and Lenz (2019) found it is possible to train computers with the company's website data if innovative and non-innovative companies are known. They used CIS data to categorizes the innovativeness of the company and neural network modeling to train it. The result was similar to the CIS, so neural network understood, what was similar to the innovative companies' webpages and it could predict if the company were innovative or not just by the text.

This paper wants to discover is it possible to use web text as data and neural network methods in the Estonian manufacturing sector to predict innovative and non-innovative indicators. There is very limited literature about studies with web scraping and neural networks, which involves companies' webpages (Gök *et.al.* 2015; Kinne and Axenberg 2018; Kinne and Lenz 2019) To the author knowledge that may be only studied on this topic in Estonia. That is why the author is testing only one sector for this study. There is a possibility to go one step further and get data, what neural network found similar to webpage texts. If similarities are known, then the author can analyze the theoretical impact of the innovation indicators on companies' innovativeness.

The first chapter explains, what measures prior studies have used and tries to answer why a new method would be useful. Mainly, the focus will be on product innovation as an innovation indicator because it is more possible that companies write about their new

product or service info, not a new machinery or equipment info, on the webpage. At the same time, it is important to review CIS data, because it is used to classify companies that have product innovation. There is also a brief review of how the neural network methods have been used before. In the second chapter, the paper describes how to scrape data on the web and prepare it for the neural network. For the analyze the author collected 2684 Estonian manufacturing companies' webpages (500 pages was the limit for the company). The chosen companies had at least 5 workers and a working webpage. Their data is filtered by the Estonian language. We can admit that larger companies have websites, and this allows us to analyze, what they have been writing about themselves. For our study, it is especially important if they write about their products and services.

This research paper could be of interest to organizations that conduct innovation research because it shows how to use new methods. Text data is very rich and has a very high dimensionality. Every research work would be welcome in this area. The author hopes that this method will give a new way to collect data, where it is no need to contact the company directly and because of that it would be a faster method. First, I want to thank my supervisors for helping me with this complicated task. They were always by my side for discussing and advising the topics of innovation, modeling and neural networks. Secondly, I want to thank my family members who helped me with emotional support and in the process of gathering data.

1. MEASUREMENT OF INNOVATION IN COMPANIES

1.1. Measurement of innovation

Throughout decades there have been attempts to measure innovation with many different indicators. Traditional indicators have been R&D expenditures and the number of patents. In many countries, R&D expenditures have been collected earlier since the 1950-s. Information about patents has been around since the 19th century (OECD 2002: 14, 200-2002; Mairesse and Mohnen 2010: 2). In this decade majority of innovation, studies have been conducted using more than just one innovation indicator. This gives us an opportunity to analyze both, quantitative and qualitative, data, not just quantitative data that has been done many years. With these data sources, economists and statisticians can more widely understand innovation and compare different countries. Different policymakers are interested in innovation studies because it shows the country growth. (Mairesse and Mohnen 2010: 2) At the beginning of the 20th century, the USA had many innovations, where the economy was in the booming stage and organizations had a lot of capital, which was a great condition to create new ideas. That is why companies found that there are other reasons for innovation too. There was a boom in innovation studies and all at once, the public wanted to know, why good ideas arise and where they come from. (Fagerberg 2013: 5-8) What are factors for innovation, which helps to create new ideas? Are there some rules? One of the reasons in Europe for the boom was an agreement to conduct CIS.

Schumpeter was one of the first people who started to find an answer to the question, is there some kind of reason for the innovation. (Scumpeter 1954: 11-14, 646, 995-996) He found some criteria must be met before innovation can begin. There must be a fundamental uncertainty and there must be a need to move faster before someone else does. Innovation does not come by itself or from luck. It forms habits and rituals, which

creates conditions for creative thinking. These conditions need to be managed. Companies do not strive for innovation because they think of the greater good of society, but because they wish to get more clients or profit from the markets (Fagerberg 2013: 11, 13-14). Thus, good management is a must. The main objective should be the decrease in research and development time. It gives the ability to lower uncertainty and motivates companies to do innovation faster than others (Fagerberg 2013: 17). In other words, companies must adapt and learn fast from their mistakes. So good innovation management is adaptable and ready to learn from mistakes.

Two types of innovation surveys have a different way to define innovation. One way the focus is on a company level and researchers ask from companies about the general innovation inputs (R&D and non-R&D inputs, like expenditures) and outputs (patents) (Nagaoka *et al.* 2010: 1085). It is defined that, innovation is a successful introduction of a new product or service to the market (Archibugi and Planta 1996: 153-154). It is an original way of defining innovation, which uses quantitative data for innovation research (like accounting indicators). Using this method innovation is defined very narrowly, but for the researchers, this way of defining innovation is objective, because it focuses on the objectively occurred technological innovation processes (Smith 2009: 161). The other way is to study important technological innovation and changes in there. It is possible through examples like new product launches or expert thoughts. This way innovation is defined as something new to the company. It can be a new process, product or service, organizational, or marketing method. (Archibugi and Planta 1996: 153-154; Mohr 1969: 112) That is why often researchers call this approach as subjective because innovation focuses is on the innovation process. An example of this would be CIS. The survey collects info about companies accounting indicators but more for the analyze the survey asks companies to answer qualitative questions and they combine both (Eurostat 2013). Although the objective way looks mainly on radical innovations then subjective way tries to consider incremental changes. Radical innovations were more important, but even Schumpeter admitted that sometimes radical innovations need incremental innovations. For example, to start mass produce airplanes or cars there was a need for special materials first (Fagerberg 2013: 8-9). So, radical innovation can be a much slower process than incremental innovation. Defining innovations only as radical may not give us enough information about changes. For some products, incremental innovation may grow to

radical innovation. Like today's world, we see a new rise of electric cars. It is because of the battery backs of the electric cars are getting better. (Toll 2017: 21-34) But the reason for the battery backs getting better lies in the discovery of new chemical combinations in small batteries. It is incremental (improvement in small batteries) but it can lead to radical innovation (wide adaptation for electric cars). So, looking into an incremental innovation is important. But answering innovation from the viewer standpoint like „is innovation something new to the world" or „new to the company" may lead to arguing and are not answered fundamentally in science. Rather researchers just define which way their work tends to go (Smith 2009: 149, 161). Indeed, studying only „new to the world" (radical innovation) may give a clearer understanding of innovation. But today's rapidly changing world every incremental change may be new innovation because competition between companies has increased. Every new change may be a good edge in competition.

Classically there has been known product innovation and service innovation. But to describe innovation with these two alone would be complicated. Oslo manual divided innovation into four types: product, process, organizational, and marketing innovation. Joseph Schumpeter defined, in the same way, five types of innovation, but it was similar to the later Oslo manual. He had a new product (1. product innovation in Oslo manual), a new method (2. process innovation), a new supplier or new market exploitation (3. marketing innovation), and organizing organization (4. organizational innovation) (OECD 2018: 70-73). Product innovation is when a company will bring new products or services to the market and it is significantly improved than previous products or services. The goal of product innovation is to maximize profit because a new product can give a competitive edge or can be just an upgrade that customers want. It should be specified for who this innovation is new to. Product innovation can be new to only the company, to a country, or globally. Process innovation is a change in the production or managing process. The main reason to do a process innovation is to reduce the price per unit (Griffith *et al.* 2006: 492). Organizational innovation is reflected in organizational internal and external relationships. Internal relationship means a change in the working techniques or physical workspace. External relationship means managing ways with partners and clients. But there should be noted that organizational innovation is not considered when some new ideas will come from the bottom of the organization, but it does not reach to the top (Griffith *et al.* 2006: 485). Marketing innovation includes the

changes in the design, packaging, layout in the shop, promotion, or in price (Oslo Manual2005: 54; Mairesse and Mohnen 2010: 7). The company is innovative if it meets one type of innovation criteria or is the middle of its process. But companies can have different innovation types at the same time. For instance, in the Netherlands research found that in the manufacturing product and process innovation are strongly correlated, while in the service sector process and organizational innovation is related. Both sectors had the same level of relation between product and organizational innovation (Polder *et al.* 2009: 17-18; Mairesse and Mohnen 2010: 22). Griffith *et al.* (2006) found that product innovation is three times more founded in companies than process innovation (Griffith *et al.* 2006: 493). It is important to remind that companies should not pick, what type of innovation they are doing, this may lead them into a narrow definition of their goal. They do not need to do innovation for the sake of innovation. Companies should let innovation be a natural and justified process (Fagerberg 2013: 13).

Generally, innovation studies are qualitative, subjective, and censored (not publicly available). Censored in the meaning is if someone wants to use new data for the innovation research it needs to have certain permission, even for the national studies. Data for these kinds of studies are not usually open for the public and if they are, then they are already old for these kinds of studies. And smaller studies what companies do are even sometimes prohibited to share, because they may have companies' secrets. So, censored has in itself a broad meaning. Innovation studies try to discover new things, so it is normal to use qualitative data. Qualitative data has more information but a higher bias. For example, if the company increased its revenue over some period about 15-20% then quantitative data may qualify it as an innovative company, because of some ratio. It depends on the sensibility of scope, is 20% enough to be innovative. But with qualitative data company must describe what they did during that period and that alone can be a sign that the company was innovative because they tried some new things first time successfully. (Mairesse and Mohnen 2010: 8) In some decades ago researchers used more qualitative data to analyze inputs of the innovation. They found that knowledge was the product of investments and innovation output comes from knowledge. (Griffith *et al.* 2006: 485) Innovation inputs are then correlated with new knowledge. But to study the creation of knowledge, it is very hard to quantify because it is describable. Quantitative

data is less informative than qualitative data, but it has less measuring bias and that makes it easier to test (answers for more “yes/no questions”).

Companies are usually divided roughly into the low- or high-tech groups. Usually, production companies are in low tech. where Sectors what are heavily labor and knowledge needed, like IT and production of hardware, are considered a high-tech and usually associated with innovation. But innovation can be found in both (Fagerberg 2013: 38). High-tech companies must look for wider and deeper knowledge to gather information for the innovation, that is why they have to look outside of their company for information. Low-tech companies may not have enough resources to look information from outside and that is why there is a possibility that sticks only with inside information (Klevorick *et al.* 1995 referred through Laursen and Salter 2006: 133). But the latter can affect the company's innovation negatively.

The most common method to measure the activities of innovation are observed with percent of R&D costs and its contribution to the overall sales. Most studies show that consistent contribution to R&D is one of the main positive indicators in innovation. (Mairesse and Mohnen 2010: 14-16; Raymond *et al*. 2006: 31). R&D costs are a regular indicator of innovation, but it shows only the input, not the output of innovation. Because this indicator is a number, it does not show what products or services were improved. (Lhuillery *et al.* 2016: 7) The company needed to have a separate R&D department and it was easier to look at that department's costs. The only criteria were that company needed to have at least one full-time employee for the researches of product and service. (Kleinknecht and Reijnen 1991; Bönte and Keilbach 2005) This raised a problem, that company must be large enough to be able to hire a person in a research position. This means R&D must be intentional. The unintentional process is not R&D and its layered as creative work and in the classical sense is not valued. This means that R&D cannot be an accidental process but is a planned or organized activity. (Godin 1983) At the same time, there is evidence that industrial R&D is usually not planned or organized activity because there is often a lack of research department for R&D or the budget for it. These reasons lead companies to sometimes declare that they do not do R&D because they think it is only related to large companies. (Chen *et al.* 2017: 697, 705-706) Companies can innovate even if they do not have a separate R&D department. More sophisticated

employees can give faster and focused development in their areas, but that does not report in R&D costs anymore. To analyze R&D, the company must account for it separately. In this decade R&D costs might be in employees' paychecks or production costs, but R&D costs cannot be overall of these costs, because some of them are just essential costs to business and do not contribute to innovation. That is why R&D costs need to be a separate system and doing so on a large scale is difficult. The result is that the big part is left out of the research of innovation and researchers have tried to fix it by lowering the standards of defining R&D (Bönte and Keilbach 2005: 298-299). So, R&D has a long history and with production, it was easier to account for it. But with the last decade changes, this indicator must be more detailed and that leads it in my view more towards to the subjective view. Of course, the classical view of R&D is objective, but in this decade, it does not capture enough of the innovation. There is a need for the new indicator, which can address smaller companies too and it all starts with getting whatever data researchers get out of them, like accounting indicators, social media reports, or even their webpage text.

Another classical objective indicator is related to patents. Patents are a public agreement between the inventor and government that gives to the inventor a monopoly position for technology (Smith 2009: 158). Patents are publicly available to everyone and that is a good thing. (*Ibid*.: 159) But they need a long time to apply, normally 36 months in the USA (Nagaoka *et al.* 2010: 1087, 1090). The researcher usually looks at this topic using indicators like the number of patents, number of patent applications, and number of citations. These indicators show the output of innovation and are considered with an objective view. Problem is that this indicator shows usually only technological improvement. Patents only show the first part of the classical innovation definition, but they do not show were inventions successful in the market or not (Smith 2009: 160). Patents are considered only for applied products and most (about 90 %) of them are just incremental improvements. Besides, patents are used in certain sectors, like in drug production or biotechnology (Fagerberg 2013: 26). So, using this indicator can lead to a very skewed understanding of innovation. In this decade there is a rise in understanding intellectual property rights and many companies prefer to use this before patents. To create an agreement to define intellectual property is cheaper and faster. (Nagaoka *et al.* 2010: 1100, 1106) Countries have collected patent information over a century and there

are now lots of data, but it involves only large companies. In this decade intellectual property has a wider range of understanding innovation then patents do, and it involves the different sizes of companies, not just the large ones.

There are more innovation indicators in partnership, and management. Even in production indicator like total factor productivity (TFT) is associated with innovation, but these are more non-standard indicators and they were created to counter weaknesses of indicators talked in this chapter. It is important to understand the history of innovation studies to move forward. It is not recommended to choose certain factors for indicators, but better is to look at first all the factors that you can collect because it can give way broader knowledge about innovation. Furthermore, some factors that are important for innovation indicators might not seem important at first but could eventually be significant. (Fagerberg 2013: 15) That's why analyzing companies' website text should be open-minded and research in this area would be essential because there is a lot to cover. It reminds us not to apply too many restrictions in the text analyses at first**.** In this chapter, the author gave an overlook of indicators, that were considered objective. In the next chapter, the author dives into one particular source of innovation indicators, a Community Innovation Survey, which is considered subjective but may give significantly more information about the innovation.

1.2. Community innovation survey and it’s uses

In 1992 OECD and Eurostat created Oslo Manual to create a standardized framework for larger innovation studies. It defined innovation in the European context and was necessary to create uniform understanding among European countries. (Godin 2002: 16-17) The company can be named innovative before the development process is finished. In many ways "new" is not clearly defined, for instance, is "new product" or "new service" considered a new for the company or a new for the whole market? So, it was necessary to lay the foundation for the innovation studies framework as it states, which qualitative data is necessary. Innovation studies based on CIS considers both the innovation inputs and outputs of the company. At the same time, it has the opportunity to explore behavioral and organizational dimensions too in addition to technological innovations. (Mairesse and

Mohnen, 2010: 6) CIS thus gave countries new ideas to study innovation, and over the years it has evolved like an innovation process itself. First, it searched only product innovation in 1992 European member states, and this decade, there are more innovation types and more countries involved.

Innovation survey that stems from the Oslo Manual and is known in Europe as Community Innovation Survey (CIS) and is conducted regularly. Since 2007 CIS has been conducted every 2 years and is named by the year when the survey was published. The latest surveys were CIS2018 and CIS2020. These surveys are actively collected and time lags are getting the data because Eurostat needs to summarize data from all European countries. That is why CIS2018 will be available in Eurostat in 30.10.2020 (CIS 2018, 2020). But CIS2020 data is actively collected from 2019 to 2020. CIS coverage in European member states has grown over the years. There were some obstacles in CIS1, the data was not harmonized, there were no clear standards and there was limited time to conduct the survey. In the beginning, only manufacturing companies were studied (CIS1), then service enterprises were added (CIS2). All the data from European countries is now harmonized and countries could ask additional questions in their surveys if they need to. (Description of the dataset, 2020; Mairesse and Mohnen, 2010: 7). This encourages some countries to also conduct their national surveys with alongside the CIS. In general, these studies are not based on the Oslo Manual and only cover a specific part of innovation. Thorough surveys of individual companies (micro surveys) are particularly difficult, as the information obtained from these surveys may not be fitting for other companies (Bogliacino *et.al*, 2009: 15-17). As CIS analysis is sometimes classified as subjective, its criticism is that innovation inputs and outputs are limited. CIS was also originally designed for industrial companies to learn about product innovation. Later other types of innovation have gradually emerged (Smith 2009: 162-163, 169). The Oslo Manual and the CIS distinguished novelty at the company, sector, and national level. There is no global or world level because innovation in CIS definition would be more similar to the imitation than innovation (Fagerberg 2013: 26). Product innovation was asked directly ("has the company launched a new product to the market?") and after that, there were specifying questions about its novelty (*Ibid*.: 165). Finally, this provided an opportunity to study the concepts of novelty and change. Many countries have started to gather only qualitative information on this topic because it is difficult or only a strategic decision for

companies to point out the costs for the innovation. In CIS, the company's management has been asked to answer a qualitative survey, and according to their information, companies are divided by engagement on innovative activities. Product innovation is easier to classify, i.e., in case of product innovation, there is some new product or service. Marketing, process, and organizational functions are evaluated through the company (Bunderson and Sutcliffe 2002: 887-889; Bogers and Lhuillery 2011: 589-590). As there are many methodologies for researching this topic, this means that it may be better to study innovation with micro-data within the sector than with macro-data across countries. Keith Smith (2009) even suggests innovation studies should avoid deep search, like confirming ground-breaking rules. Innovation studies should first be discovering, broad looking with new methods. It was noted that CIS works even better with manufacturing data than with the service sector data (Smith 2009: 169). So, to test a new method, it is easier for the author to analyze manufacturing websites first and see, what the results of that will be.

Many measures, no matter whether quantitative or qualitative, in innovation research are subjective because they depend on the assessor. A good example is an increase in sales revenue for the new product because it is usually rounded (like 10, 15 or 20 percent). This measure could also be viewed as a categorical identifier if it is defined in ranges. Although companies categorize their products and it is usually possible to get exact numbers, it is much more difficult to understand, how a new product affects sales of other products. It is also questionable what is new for the company and what is new for the market. Oslo Manual states that if there is no improvement in product or service and if the latest trend of design is not captured, then the company is less innovative (Mairesse and Mohnen 2010: 8-9). Matching a trend gives companies a possibility to rank innovation, but that is a trap because then there is a possibility to learn only the most innovative companies. Other companies will be set in the background and most of the cases "other companies" are the smaller ones. So, ranking in innovation is not a way to study novelty.

The other quantitative argument is about research and development expenditures and what should be considered as such funds as we discussed in the previous chapter. Some consider that this is a cost that the company has spent on separate development research. The Oslo manual states that, in addition to research and development costs, these are also

the costs of staff training, engineering costs, design, and marketing costs. These are all costs related to the company's input. There are also subjective cost savings, that can come from process innovation. It may be difficult to get the latter information from the companies as it is often related to a trade secret, where it is not recommended to share this information (Mairesse and Mohnen 2010: 14). That is why getting this information is up to the company representative to answer. Even though trade secret information does not have to be answered, at least CIS asks that, and if it is answered, CIS analyses it.

If innovation is viewed over time it must be noted that it is a dynamic phenomenon (Griffith *et.al.* 2006: 493). For example, a company can be innovative at one time but in the next period, it is not. To be an innovation is a constant improvement. The problem with CIS input data is that companies' information is collected partly over sectors. To compensate this the sample is rotating in every CIS. It gives better coverage of the market over the years, but it brakes the opportunity to study a time series with CIS. So, to understand innovation there is a need to study companies over time (Laursen and Salter, 2006: 147). Some countries like Germany have noted that and they are conducting much broader innovation studies around CIS. The Mannheim Innovation Panel (MIP) conducts innovation surveys every year. These surveys are based on CIS and they are conducting it the same way as CIS. Germany reports about CIS to Eurostat like other countries, every two years. But they have more continuous data about companies and their innovation. (Gault2013: 143-144) This gives more information to understand innovation in a certain country and defines companies' innovation from its definition, an innovative company is a company that searches constant improvements.

In innovation research, the subjectivity of the observer needs to be justified more, because sometimes the interpretation depends on the respondent (his background, culture, and the situation of the moment) (Fagerberg 2013: 26). Not only does the view not necessarily help the company, but the market must also be understood. Sometimes it is difficult to access data because usually it must be collected at the company level. Also, it might be difficult to access databases. Another difficulty arises from the confidentiality and anonymity of participants. Ensuring anonymity complicates research and sometimes even interpretation (Mairesse and Mohnen 2010: 25-26). Therefore, the information on the

websites can be good qualitative information in machine learning, as it uses the information available to everyone and the companies in the whole market.

Innovation theory supports using new methods, like online surveys, e-administrative records, data scraping, or social media research (Sauermann and Roarch 2013: 274; Geuna *et. al.* 2015: 1656). It is not necessary to do it with direct objective data. The researcher can use many indirect methods to analyze innovation in companies because the meaning of innovation has changed over the decades. In this rapidly changing time companies are looking for even the smallest incremental improvement to improve their competitiveness and for that, we need to understand, what are the bases for change.

1.3. Neural network as a new method in innovation studies

Technology has made it possible to share a great number of digital texts. This allows for collecting it and analyzing it. In social science words and text have a lot of meaning. Decoding a text is an opportunity to get far richer explanations for phenomena than more structured kinds of data. In this decade there has been a rise in the empirical economics studies that uses text as data (Gentzkow *et. al.* 2019; 535). In macroeconomics, there have been different studies with neural networks: unemployment, inflation. (Wanto *et al*. 2018; Choudhary 2012) In marketing, there have been studies with neural networks to understand the drivers for consumer decision making. (Baesens *et al*. 2002) In political economy has been studied text from politicians. (Rao, Spasojevic 2016) As a result, web scraping and text mining have begun to rise in economic studies as a novel tool and insights to economists. (Levenberg *et al.* 2014: 109) It should be noted that a neural network is a group of methods, not just a single selection. In the natural language processing (NLP) has been great improvement as the increase of the computational power has allowed giving a new methodological approach. (Collobert, Weston 2008: 161). This all has given the option to use text data in neural networks. In-text documents classification these methods have given a promising result. (Kim 2014: 1751; Yang *et al*. 2016: 1487).

There are strongpoints in using text as data than traditional innovation indicators from questionnaire-based surveys or patient-based studies. (Nagaoka *et al*. 2010: 1085–1106) Main points to use text as data are coverage, granularity, timeliness, and cost. (Kinne, Axenbeck 2018: 2) Firstly coverage with text as data is all about how to get them. Most of the data is free and open on the internet, for example, news articles or product reviews. Even on open databases, there is a lot of information that could be used with texts. There is no need to contact someone or think about how much data should be covered to describe a phenomenon. Usually these kinds of studies all the total sample is represented. Secondly, granularity because the text is very rich in information. It indeed depends on the methods of what to use, but overall, it allows looking deeper meanings of the different phenomenon or behind the meanings that we do not know thoroughly yet, like innovation and novelty. Thirdly timeliness can be divided by the time how much time it takes to get a total sample or how much data can be mining in online (reducing time lag between gathering data and analyzing it with some phenomenon). In some areas, real-time data that could be collected automatically can make a huge difference in performance. Fourthly cost in time and money is usually cheaper that interview or question-based studies where people must get data from other people. These points are very attractive and convincing to use in economic studies, but every good thing has its bad aspects too. Because it is a new method and has a lot of opportunities there are no hands-on beginner-friendly programs yet. To understand machine learning researchers must know about programming languages, mainly in Python or R. There are not good universal hands-on programs to use with these kinds of methods. But even if the researcher knows programming languages there are other problems too.

The main problem with analyzing text with 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 predicted values of unknown outcomes. Later the outcome is in subsequent descriptive or causal analyses (Gentzkow *et. al.* 2019: 536). So, if the problem/phenomenon needs a lot of input data then it needs a lot of computational power too. There are simple methods that look only how many words are used in data and map it out. For better results, there are methods that look nearby words too, but it multiplies the opportunities for how many words and combinations are in data and it complicates calculations.

Overall innovation studies are more of the novel type of studies just like studies with machine learning and neural networks. It is easier to train neural networks when there are some sentiments or classifications with texts. CIS what has been conducted many years already, has classified companies already into the innovative and non-innovative ones. CIS even separated product innovation as one of the components of overall innovation. This gives the opportunity to learn innovation in the text data and that is the focus for the next chapter.

2. PREDICTING INNOVATIVE MANUFACTURING COMPANIES WEBPAGES IN ESTONIA

2.1. Data collection process

Most of the companies use their webpages as one of the main mediums, where they can publish info about their products and processes (Gök *et.al.* 2015: 654). With great probability, companies put information about their new products on to their webpages, because it has marketing value. Therefore, product innovation should be more probably revealed in companies' webpages than process innovation. This website info is the data for machine learning available as freeware and there are many libraries free to use. In this chapter's main goal is to describe how to do a product innovation predicting model which could predict what companies have product innovation and which do not.

There have been two German studies from Jan Kinne and David Lenz 2019 and Jan Kinne and Janna Axenbeck 2018. Both of them used in Germany the MUP (The Mannheim Enterprise Panel) database, which is similar to the Estonian Business Register (est *Äriregister*). Because MUP collects data about companies, they have access to the German companies' websites list. With this list, it is possible for my website data. The last part was the input data for the neural network to learn. In this test group, they had innovative and non-innovative companies together. The main goal for the neural network was to understand, what was in common for webpages used by innovative companies. In Germany, there were 2,52 million companies and 1,15 million had webpages, so roughly 46 % (in both studies). Companies that had less than 5 workers and not active before 2018 (study conducted 2019) were removed (Kinne and Axenbeck, 2018: 9-12; Kinne and Lenz, 2019: 2-3, 6). Kinne and Lenz tried to find in particular how to predict product innovation with companies' website data because this was easier to conduct with their

methods (Kinne and Lenz 2019: 5). Both German studies did not comment on marketing innovation though it can be a possibility to learn too with CIS. Kinne and Lenz combined selected MUP and MIP (The Mannheim Innovation Panel) data. MIP in Germany is similar to Estonian CIS survey run by Statistics Estonia (in Estonian *Statistikaamet*) database. MIP is doing every year innovation survey, so they used 3 years (2015-2017) of innovation data from Germany to show which companies are innovative and which not. There is a time lag between scraping website data and determining innovation in CIS. Like in Germans studies website scraping was done in 2018 but CIS data was from 2015-2017 (they had MIP survey data, what was conducted every year). They filtered out companies who were innovative or non-innovative 3 years in the row. The reasoning behind it was that innovative status is dynamic over time and it makes a clearer border for neural network training. At the same time, this meant that the company had to be at least 3 years old (Kinne and Lenz 2019: 2-3). The author of this paper will compare the early 2020 website data with 2015-2016 and 2017-2018 CIS data in Estonia (Ettevõtete innovatsiooniuuring..., 2020). The reasoning behind using 2 sets of CIS data is that most of the manufacturing companies were the same. There was a need to get more companies for learning.

The website data was collected from 2019 December till 2020 May. All the processes were separated into seven stages (look Figure 1) because it contained different programs and scripts. Firstly, the main goal for the author was to get company URLs (Uniform Resource Locator). For that he had two options: find information himself or cooperate with databases. Because the author needed data about the number of workers and the company registration date then he opted to collect it himself. The author also wanted to learn more about the process of web scraping and prove that this process can be done freely at home. For the web scraping author used Estonian companies' database named teatmik.ee, because it was fell structured for html scraping. Also, they collect data from seven different sources. The first stage was done in two parts. The first part was to get a list of companies' names and register numbers to get their database URL. For example, if we want to find info about company MasterResearchPaper OÜ with register number 12345678, the URL for that would be https://www.teatmik.ee/et/personlegal/12345678-MasterResearchPaper-OU. So, URL has the main part https://www.teatmik.ee/et/personlegal/ plus register number plus name. In the name part,

all spaces were converted to hyphens and Estonian letters (ü=u, õ=o, ö=o, ä=a) were converted to ASCII letters. The second part was to collect companies' metadata with a list of first part URLs. The final result from the first stage was in Microsoft Excel file with all the manufacturing companies (17 647 companies). Data was gathered by counties in Estonia. The author collected all the data about companies' websites with 5 computers (1 desktop, 2 laptops, and 2 Raspberry Pi-s). There are companies, who have classified themselves in different sectors and their main one is not manufacturing. That is why there are some wholesalers, who produce only one product, but they classify themselves as manufacturing companies too.

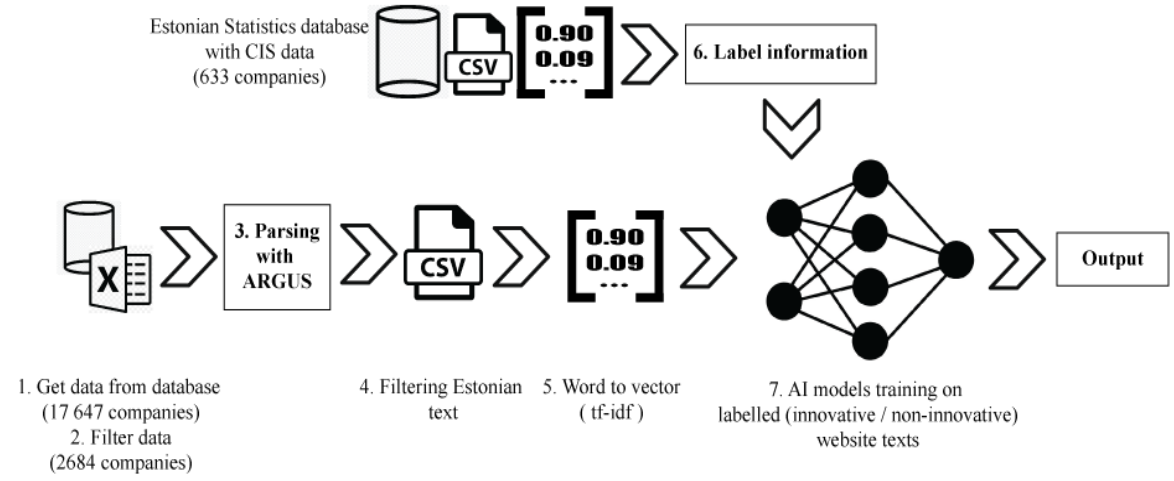


Figure 1. Product innovation prediction model.
Source: (Kinne and Lenz, 2019: 2; compiled by the author)

In the second stage, the author had to filter out companies who met conditions: working webpage and at least 5 workers in 2019. Only 2684 companies were met the criteria. Analyses with them are done in the next chapter. For extra background data languages of the webpages were collected by going on the manufacturing companies' webpages and by looking at what languages they had. Outcomes were recorded in the Microsoft Excel sheet.

In the third stage, the author used a Python web scraping program that was used in both German studies. It is called Automated Robot for Generic Universal Scraping (ARGUS) (Kinne and Lenz, 2019: 3; Kinne and Axenbeck, 2018: 9-11; ARGUS: Automated..., 2020). With ARGUS, it is possible to scrape all web page texts and its links (on-site and

external links). Before the scraping user must set, how many pages ARGUS should scan (look appendix 1). Scanning will be done with structural order, where firstly scanned pages are in the main menu and next pages in the submenu and so on. Kinne and Lenz argued that the 250 pages limit should be covering 90% of the companies, but they put 100 pages for their analysis. Idea was that company information data (like "About us") should be in 100 pages, later there should only be product or service information. A page limit of 500 was used because to understand product innovation. So, it is important to know that later text too. With 100 pages, scanning and filtering data should be faster and because Germany has a lot more companies than in Estonia, there is a possibility that they would not need too much data on the webpages. Kinne and Lenz (2019) had 8080 companies all over Germany, but they were well selected (over the three years) (Kinne and Lenz, 2019: 3). A big difference with this paper and Germans studies is that in this paper only the manufacturing sector was tested (2684 companies). That is why the author went deeper with website data, so it took more time to filter. ARGUS output is given in CSV-format (Comma-Separated Values format).

In the fourth stage, Kinne and Lenz used only German language and filtered out other languages. Filtering was done by ARGUS. (Kinne and Lenz, 2019: 3) With the Estonian language, there were not working options, and the author filtered Estonian data by hand about 4 months (February 2020 till the end of May 2020). Text filtering was another reason, why only the manufacturing sector was focused on because many of the companies in Estonia use English websites too for the export. For this paper, only Estonian data was used, but because there was a need to check if websites working, then data about webpage languages were collected too. Languages were Estonian (EST), English (ENG), or both (EST/ENG). Other languages were filtered out and marked only with number how many languages were on the page. English was the most dominant language from other non-native languages. About 64% used English with Estonian language and 10% used only English or another language. Look analyses with manufacturing companies' website languages in appendix 2. Filtered Estonian text was saved in txt-format files (like containers) because Excel has a cell limit of 32 767 characters (MS Excel support, 2020). Later all the text was gathered into a one CSV-format fail where companies were indexed by the order number.

The fifth stage is a step before analyzing and training could begin. This step's goal is to convert text to understandable for the machine. For that text was converted into a word vectors in lowercase. Kinne and Lenz did this task in their research too (Kinne and Lenz, 2019: 3). Lowercase helps to keep the word count lower and equalizes words, what computer may think are different words. Like the words "paper" and "Paper" are two different words for the computer. Next, raw data was cleaned from elements that do not tell anything, such as punctuation, numbers (dates were not needed), HTML tags, and so on. Finally, there is a need to filter out words that are too common called "stopwords". Common words such as "I", "we", "how" have very high frequency but they do not have value, because they do not tell the meaning of the phenomenon what is being searched. To lower the vocabulary, count the text was lemmatized. It is one of the natural language processes (Gentzkow et. al., 2019; 538). Meaning every word got its base form, the example in Estonian "toodetes" would convert into the word "toode". Later word vectors were converted into a "term frequency-inverse document frequency" vector (shortly tf-idf). It is a matrix that is intended to reflect how important a word is to a document in a corpus. Term frequency shows how many times a word is in the one document. But some words have a high frequency in every document (over the corpus) that is why usually it is needed to balance them. So inverse document frequency is a logarithmic part over a share of documents containing a word. Very rare words will have low tf-idf scores because they do not occur much (low tf-score). Very common words that occur most of the documents have low tf-idf score because idf-score is low. We are interested in words what have a high frequency in certain documents (like innovative and non-innovative text) because it allows us to understand a segment. If a word would be in very high frequency in every document then it does not separate a segment, like an innovative company, because it would be in non-innovative company text too. And because values are high this word would be important, and it would disrupt a calculation for the neural network. (Gentzkow et. al., 2019; 538)

The sixth step is to collect data from Statistics Estonia servers about CIS 2015-2016 and 2017-2018 results. These results will be an important input to know, which companies are innovative and which not. CIS data was filtered by the company register number, to hide company names and provide anonymity. From the CIS data product, innovators were labeled out and applied to the training set. Because all the manufacturing companies' data

were gathered before it was just a "label the company" step. Product innovation was selected because there was evidence (like product innovation was three times more founded in companies than process innovation) that it is the easiest type of innovation to define and test with (Griffith *et al.* 2006: 493). By this far it is fair to say the current paper uses CIS data, which is subjective and website text, which is the same, so overall this study is more towards the subjective view, as explained in the first chapter. So product innovation for the author is defined than just as something new for the company. The final part is to validate the model and train it to predict product innovation. The seventh step is explained in the next chapter.

2.2. Data analyses

This chapter is divided into firstly, describing background data and secondly, how the model is validated and trained. This paper looks only for one sector (manufacturing) and that is why it is good to point out data that may not be in the model training. These are the objective numbers that came with this study and are good to present too.

There were 17 647 manufacturing companies in Estonia in January 2020. 6480 of them had webpages of which was 37%. There were 2684 companies, who had a working webpage and more than five employees, which was 15% of overall companies (look Appendix 2). It was necessary to filter out inactive companies to make Estonian data comparable to the used in German studies. The summary of the number of economically active companies in the manufacturing industry is in appendix 3. The activity is defined if the company had at least 1 euro of revenue or any employees in 2019. Next, inactive companies were filtered out to get a better picture of the manufacturing sector in Estonia. Table 1 shows, how many companies in manufacturing are active. The author then calculated new percentages with the active companies to see the real usage of webpages. Figure 2 is a histogram for a visual look of the data, where „Don't have URL" is overall an active number of manufacturing companies, „URL" is how many of them have a webpage and „↑5 employees, URL" means the number of companies who have more than 5 employees and a working website. After filtering out active companies, the results of the manufacturing companies, who have a webpage is still similar to Kinne and Axenberg

paper (within 60-75%) and higher than Kinne and Lenz paper (46% overall coverage). Interestingly, both papers looked at the overall number of webpages and some companies may have 2 or more webpages (Kinne and Lenz, 2019: 2, Kinne and Axenberg, 2018: 12-13). Companies were filtered by their register number, so if one company had two or more webpages, it still counted as one company. On average 2/3 of Estonia manufacturing companies have a webpage. For the innovation analysis author used 31% of the active companies, because the author wanted to focus on the same set of companies like Kinne and Lenz did: companies who have 5 or more workers and a working webpage with the only difference, where Kinne and Lenz looked overall market and this paper looked only manufacturings sector.

Table 1. Economically active manufacturing companies and companies with more than 1 employee in 2019.

County	Nr of active companies	Companies who have a website	%	Companies who have 5 or more employees and a website	%
Valgamaa	88	82	93%	33	38%
Võrumaa	184	130	71%	56	30%
Hiiumaa	54	40	74%	13	24%
Jõgevamaa	172	106	62%	46	27%
Lääne-Virumaa	329	233	71%	98	30%
Põlvamaa	135	103	76%	37	27%
Saaremaa	221	192	87%	78	35%
Tartumaa	1178	770	65%	298	25%
Pärnumaa	341	292	86%	118	35%
Ida-Virumaa	510	290	57%	146	29%
Viljandimaa	204	137	67%	63	31%
Raplamaa	264	172	65%	67	25%
Läänemaa	151	91	60%	37	25%
Järvamaa	186	120	65%	58	31%
Harjumaa	4549	3722	82%	1536	34%
Sum	**8566**	**6480**	**76%**	**2684**	**31%**

Source: compiled by the author

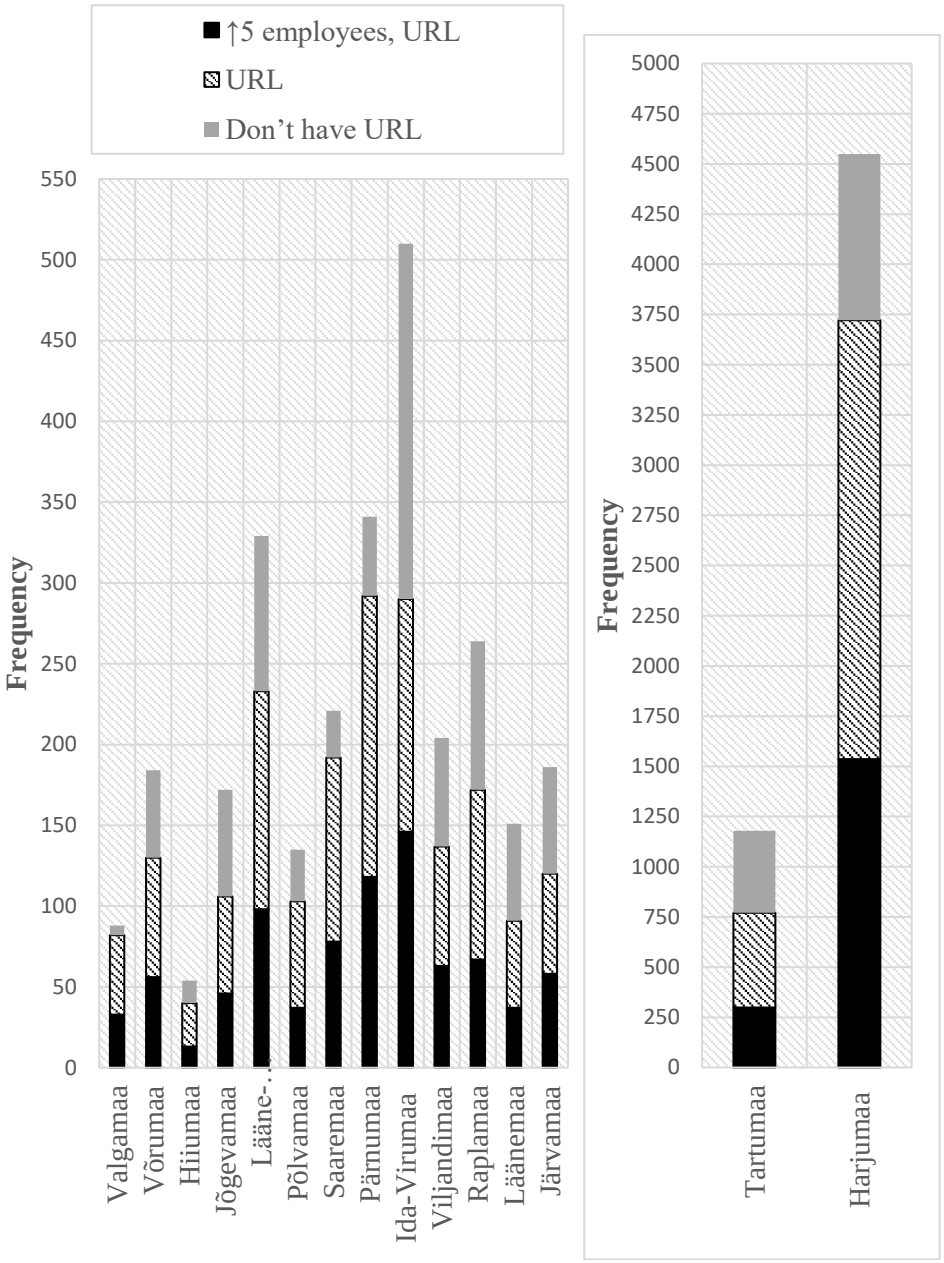


Figure 2. Visual presentation of the data of Table 2.
Source: compiled by the author

As described in step 2 in the last chapter, companies author checked what languages were used on the homepage. For the neural network testing author selected the Estonian language. Collective info about languages filtered by county is in Table 2. It shows, how many companies had only Estonian (EST), only English (ENG), or both languages (EST/ENG) on their websites. It must be noted, that if the company had only English, that does not mean it did not have other languages aside from Estonian there. For example, only English means there could be English and Finnish and only Estonian means there could be Estonian and Russian languages there. The majority had the Estonian language

on their webpages, specifically 2418 companies, which is 90% of all pages. English was used by 1710 companies, which is about 64% of manufacturing companies. It confirms that over half of the manufacturing companies use English websites in Estonia. Many companies in Estonia use English as their second language because it helps them with export markets. Overview of statistics is in Table 3.

Table 2. Estonian manufacturing companies websites by language (2020 January)

County	EST	ENG	EST/ENG	Sum
Valgamaa	8	2	23	**33**
Võrumaa	22	1	33	**56**
Hiiumaa	3	2	8	**13**
Jõgevamaa	23	5	18	**46**
Lääne-Virumaa	39	5	54	**98**
Põlvamaa	16	2	19	**37**
Saaremaa	32	8	38	**78**
Tartumaa	177	12	109	**298**
Pärnumaa	27	22	69	**118**
Ida-Virumaa	54	16	76	**146**
Viljandimaa	21	7	35	**63**
Raplamaa	33	6	28	**67**
Läänemaa	18	1	18	**37**
Järvamaa	26	2	30	**58**
Harjumaa	475	175	886	**1536**
Sum	**974**	**266**	**1444**	**2684**

Source: compiled by the author

Table 3. Estonian manufacturing companies' websites statistics about their languages. (2020 January)

Indicator	EST	ENG	EST/ENG	Overall
Average	1,27	2,09	3,84	2,73
Median	1	1	3	2
Min	1	1	1	1
Max	15	50	183	183
Standart deviation	0.77	3.42	6.54	5.08

Source: compiled by the author

Overall average language count was 2,73. That means on average every webpage had 2-3 languages in it. The minimum number of languages was naturally 1 and the maximum number of languages was 183. If the company had Estonian and English languages on the website, there is a high possibility that it had even more languages, because the average language count was 3,84 and the median was 3. The author thinks the reason behind it is that if the company exports to the other country it translates its website to that country

language too. So, combinations of languages from neighboring countries would be Estonian/English/Finnish, Estonian/English/Russian, Estonian/English/ Latvian, Estonian/English/Swedish.

From CIS 2018 and CIS 2016 combined author got 678 manufacturing companies with product innovation labels. Companies who had too little text or were under construction at a time of scanning and in July 2020 were filtered out. Some companies had only one webpage and it was their contact only and they had an innovative label. So, after filtering 633 companies remained. Table 4. show that data distribution with CIS labels was balanced. There were 309 manufacturing companies with product innovative labels and 324 non-innovative companies with product innovation.

Table 4. Manufacturing companies distribution within combined CIS2018 and CIS2016.

Description	Nr of companies	Distribution
Nr of Innovative companies (with label 1)	309	48%
Nr of Non-Innovative companies (with label 0)	324	51%
The overall number of companies	633	

Source: compiled by the author

Standart statistics about webpage text distribution is shown in the Table 5. Words distribution between the companies shows Figure 3. Vocabulary length is good to know because it sets limits for the model training computational power. Average words are good to know if embedding would be something else than tf-idf, because it shows where to do padding if needed. In tf-idf the matrix is all the companies (633) times whole vocabulary (214555). For the testing vocabulary was limited in tf-idf vector. 3 modes were used: „None“- no limit to the vocabulary, „min_df=5" - was at least five same words all over the corpus and „min_df=5,max_df=10000" - requirement was at least five same words all over the corpus and higher limit was 10000 for the same word in the corpus. With upper and lower limits corpus size was reduced to 20875. With this computational power was lowered and results with validation tests were better.

Next stratified k-fold cross valuations were applied for the model with k values 3, 4, 5, 9, and 10. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Parameter k refers to the number of the groups data sample is split into. Then the groups will be tested one by one and put back at the sample. Other groups at the same time will be training data. So, in the end, every group will be

test data one time and in other times a training data. Stratified means that the sample is split in balance. That 1 (innovative) and 0 (non-innovative) would be in the same ration in groups.

Table 5. Standart statistics over webpages texts.

Description	Statistic
Population	633
Vocabulary length	214555
Average nr of words in files	9275
Median nr of words in files	1807
The standard deviation of the number of words	19560

Source: compiled by the author

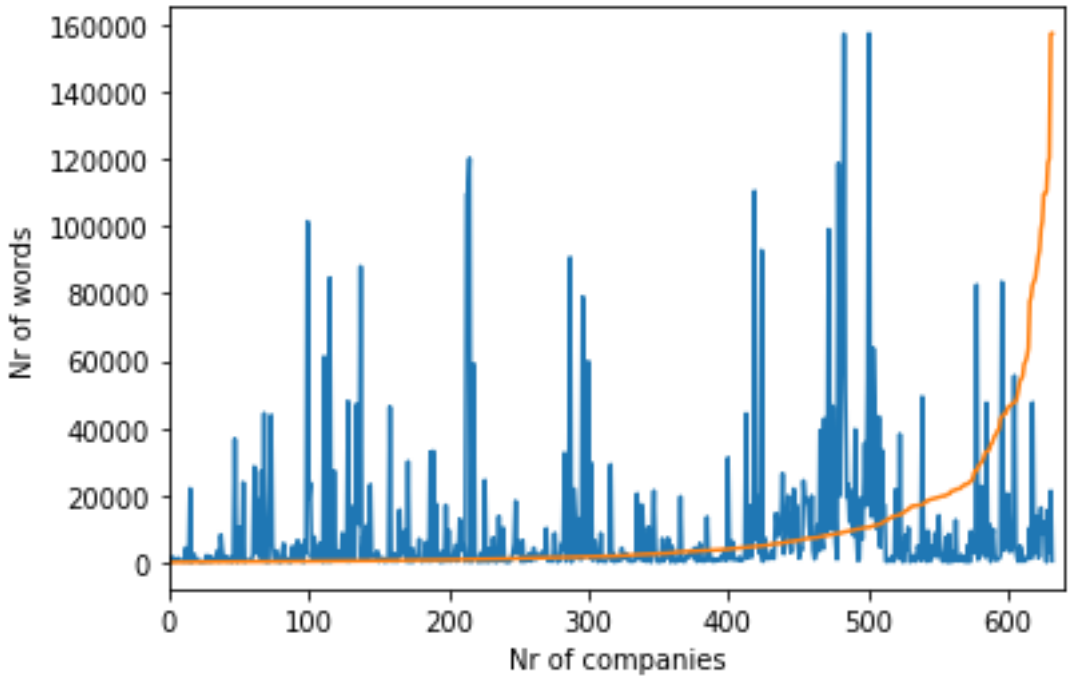


Figure 3. Number of words used in webpages (Upper is a histogram and bottom a line chart)
Source: compiled by the author

Traditional models were running to validate the model. These models were logistic regression, Bernoulli, multinomial, Gaussian, SVM (Support vector machine), random forest, k-neighbors, Gradient Boosting, Decision Tree, XG Boost. Regression methods estimate the conditional outcome distribution.

All the validated models' metrics with different tf-idf vectors and k-fold parameters were reported in Excel-file. The first look was in the accuracies to find where are the higher

ones. AUC ROC (Area Under the Curve - Receiver Operating Characteristics) analysis was conducted with the combinations of the best accuracies' models. Purpose of this analyses is to find out best validation model what would be a baseline for the later. AUC ROC is a probability curve and it is one of the most important evaluation metrics for checking any classification model's performance. It describes model capability to distinguish classes and it can be applied to different models at the same time. Metrics for these analyses are in next chapter where the models are compared. The aim of this is to find important words what were used to evaluate data accuracy. With these words it is possible to describe manufacturing companies' product innovativeness because the model has used these words highly to predict its outcome.

2.3. Results

To go on, the first look has to be at the metrics of the models, specifically accuracies. Table 6 shows the overall accuracies of the models. The accuracy value area was about 0,55-0,65. Best accuracies were on the model then it had limited vocabulary where minimum words count was 5 and a maximum 10 000. Promising results were in the forest classification with estimators 60. Also, Gradient Boosting gave the same result. Classical models Bernoulli, Multinomial, Gaussian gave the worst results compare to other models.

For better comparing AUC ROC scores were calculated and evaluated. Best results are in figure 4, where configuration, only 5 word limit to vocabulary, has the highest scores of all the AUC ROC. The best classification separation was done with model random forest with the n-estimator level at 40 because it had the highest AUC ROC score of 0,628. Right after that was the GX Boost model with AUC ROC score 0,622. Two models were picked to compare their words of importance. Metrics reports were used on both models, look Table 7.

Table 6. Overall accuracies on the models with different input data configurations (tf-idf and k-Fold).

	No Restrictions on vocabulary				Less than 5 words in the whole vocabulary					Less than 5 and more than 10000 words in the whole vocabulary are 0				
TfidfVectorizer	None				min_df=5					min_df=5,max_df=10000				
Stratified K-Fold Cross validation (k=)	3	4	5	10	3	4	5	9	10	3	4	5	9	10
LogisticRegression	0,60	0,61	0,60	0,59	0,60	0,61	0,60	0,59	0,59	0,60	0,61	0,60	0,61	0,59
SVC / SVM	0,58	0,58	0,58	0,58	0,60	0,61	0,59	0,60	0,59	0,60	0,61	0,59	0,60	0,59
BernoulliNB	0,58	0,59	0,59	0,58	0,61	0,59	0,605	0,59	0,59	0,61	0,59	0,60	0,59	0,59
MultinomialNB	0,580	0,560	0,576	0,589	0,600	0,586	0,600	0,608	0,597	0,600	0,586	0,600	0,608	0,597
GaussianNB	0,508	0,540	0,549	0,560	0,508	0,540	0,549	0,538	0,560	0,508	0,540	0,549	0,538	0,560
RandomForestClassifier														
n_estimators = 15	0,58	0,58	0,59	0,59	0,61	0,63	0,61	0,60	0,56	0,60	0,58	0,59	0,59	0,58
n_estimators = 40	0,61	0,60	0,60	0,61	0,59	0,63	0,60	0,60	0,60	0,61	0,60	0,62	0,60	0,60
n_estimators = 60	0,58	0,61	0,61	0,62	0,61	0,62	0,59	0,60	0,61	0,60	0,64	0,63	0,60	0,63
n_estimators = 100	0,60	0,62	0,64	0,61	0,61	0,61	0,59	0,59	0,61	0,62	0,62	0,60	0,62	0,62
n_estimators = 200	0,61	0,61	0,62	0,61	0,59	0,61	0,61	0,61	0,62	0,61	0,62	0,61	0,60	0,62
n_estimators = 500	0,60	0,62	0,61	0,61	0,60	0,61	0,62	0,61	0,62	0,60	0,62	0,60	0,61	0,62

Source: compiled by the author

Table 6 will continue next page (34)

Continuing in Table 6. Overall accuracies on the models with different input data configurations (tf-idf and k-Fold).

TfidfVectorizer	No Restrictions on vocabulary None				Less than 5 words in the whole vocabulary min_df=5					Less than 5 and more than 10000 words in the whole vocabulary are 0 min_df=5,max_df=10000				
Stratified K-Fold Cross validation (k=)	3	4	5	10	3	4	5	9	10	3	4	5	9	10
SVM	0,58	0,58	0,58	0,58	0,60	0,61	0,59	0,60	0,59	0,60	0,61	0,59	0,60	0,59
KNeighborsClassifier	0,56	0,54	0,55	0,52	0,55	0,54	0,55	0,53	0,53	0,55	0,54	0,55	0,53	0,53
GradientBoostingClassifier	0,60	0,61	0,60	0,61	0,58	0,62	0,61	0,59	0,61	0,59	0,63	0,61	0,59	0,63
DecisionTreeClassifier	0,55-0,58	0,54-0,56	0,549-0,57	0,53-0,57	0,52-0,56	0,55-0,58	0,56-0,58	0,54-0,55	0,53-0,55	0,54-0,56	0,56-0,60	0,56-0,57	0,55-0,58	0,53-0,54
XGBClassifier	0,58	0,61	0,57	0,60	0,60	0,62	0,59	0,61	0,60	0,60	0,62	0,59	0,61	0,60

Source: compiled by the author

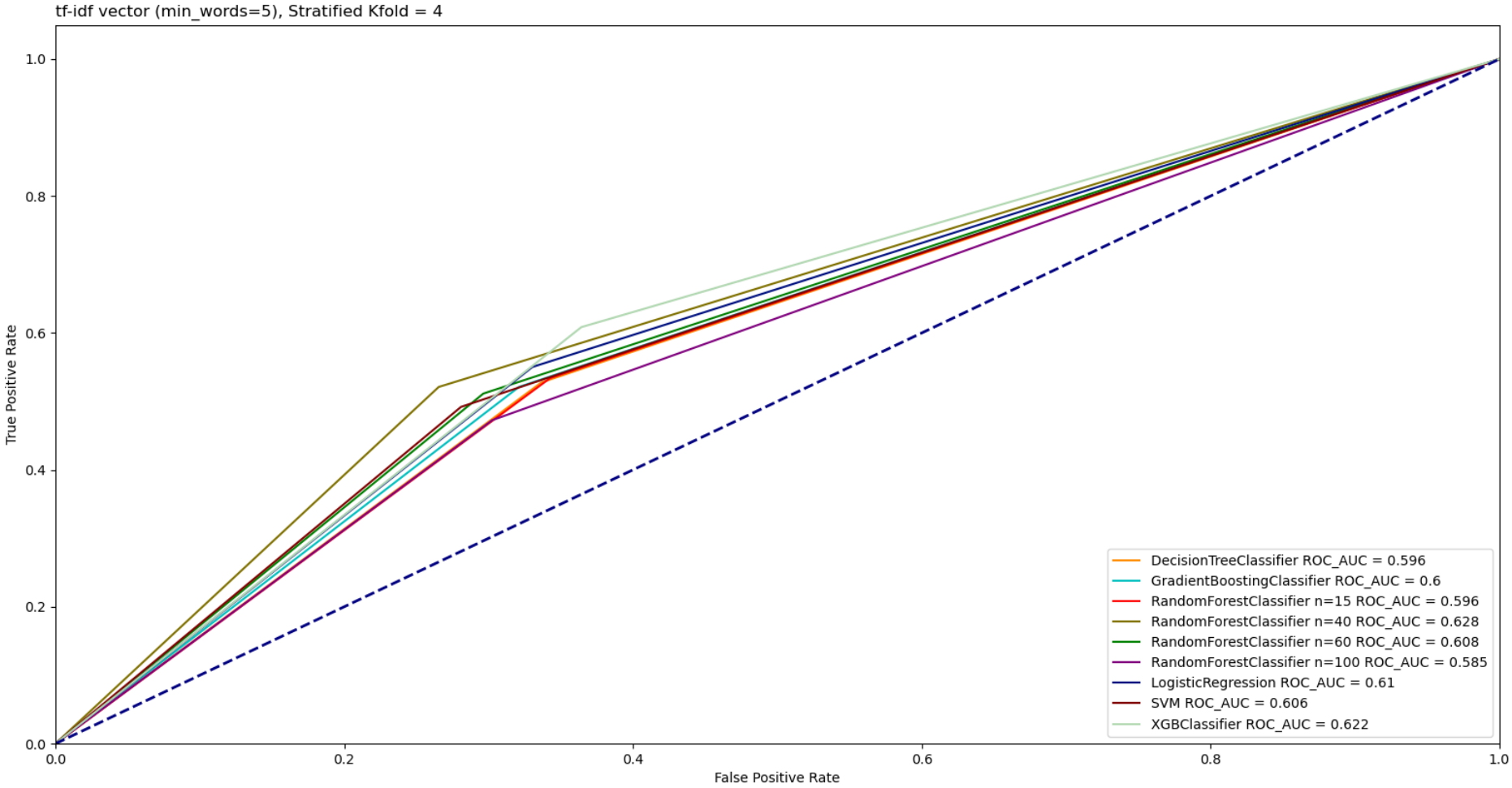


Figure 4. AUC ROC analyses in a line plot.
Source: compiled by the author

Table 7. RandomForestClassifier output with n_estimator 40, tf-idf limits (min 5 words), and k=4.

Accuracy: 0.61%
Classification_report

	precision	recall	f1-score	support
0	0.60	0.71	0.65	324
1	0.63	0.51	0.56	309
Accuracy			0.61	633
macro avg	0.62	0.61	0.61	633
weighted avg	0.61	0.61	0.61	633

Crosstab

col_0	**0**	**1**
row_0		
0	229	95
1	150	159

Source: compiled by the author

Table 8. GX Boost output with tf-idf limits (min 5 words and max 10000 words) and k=4.

Accuracy: 0.60%
Classification_report

	precision	recall	f1-score	support
0	0.60	0.63	0.61	324
1	0.59	0.57	0.58	309
accuracy			0.6	633
macro avg	0.6	0.6	0.6	633
weighted avg	0.6	0.6	0.6	633

Crosstab

col_0	**0**	**1**
row_0		
0	203	121
1	134	175

Source: compiled by the author

Autor used both models to find a figure of importance plots and compared these plots with each other to find what words are important in both models (look Figures 5 and 6). This method is normally used to understand what indicators have higher weight (importance) in the model. But with the tf-idf matrix, every word is an indicator. That's

why looking at the top 50 words and compare them would give some idea about the use of website data. Idea is that if a word has a high importance rating in both models then this word has significance in data and could be interpreted with Estonian manufacturing sector product innovation. Words are categorized into different topics and then explained why it is like that.

Table 8. has an overview of the categories of the top 50 words in the best-validated models. There were 3 categories of words in both models: product characteristics, places, and time. Words that could be interpreted with product characteristics are "price", "freshness", and "tasty". It indicates that some of the innovative companies may be food producers. They might have a big webpage with lots of text describing food. Interesting was the second category named "places". Because these are the export countries where Estonian manufacturing companies are mainly exporting in Europe. One of the biggest export countries is Finland and another in Italy. These countries have innovation centers in them, and Estonian manufacturing companies may be contributing to them (mainly Finland). The effect can be the opposite that Estonian companies are importing their products too (if we look at Italy), but it still contributes to Estonia in the product innovation. Tallinn is the capital of Estonia and most of the manufacturing companies are located there (look table 1 "Harjumaa label"). Because more manufacturing companies are there then the number of innovating companies is also higher there. Last in this category is "Põldsamaa". It is a city and a food-producing company's name at the same time. So, this company may be innovative even without looking at innovativeness in its label. In the time category interesting word is "märts" which means in English a month in March. Because the 2020 coronavirus pandemic hit Estonia in March, it might be involved why there is a month of march on top of both models. Companies webpage scanning was done in February till May 2020 so if other months would be important then they would have been in this top too, but only a month of March is in models. Innovative companies may have reported the news at that time, and they may have used to describe that month.

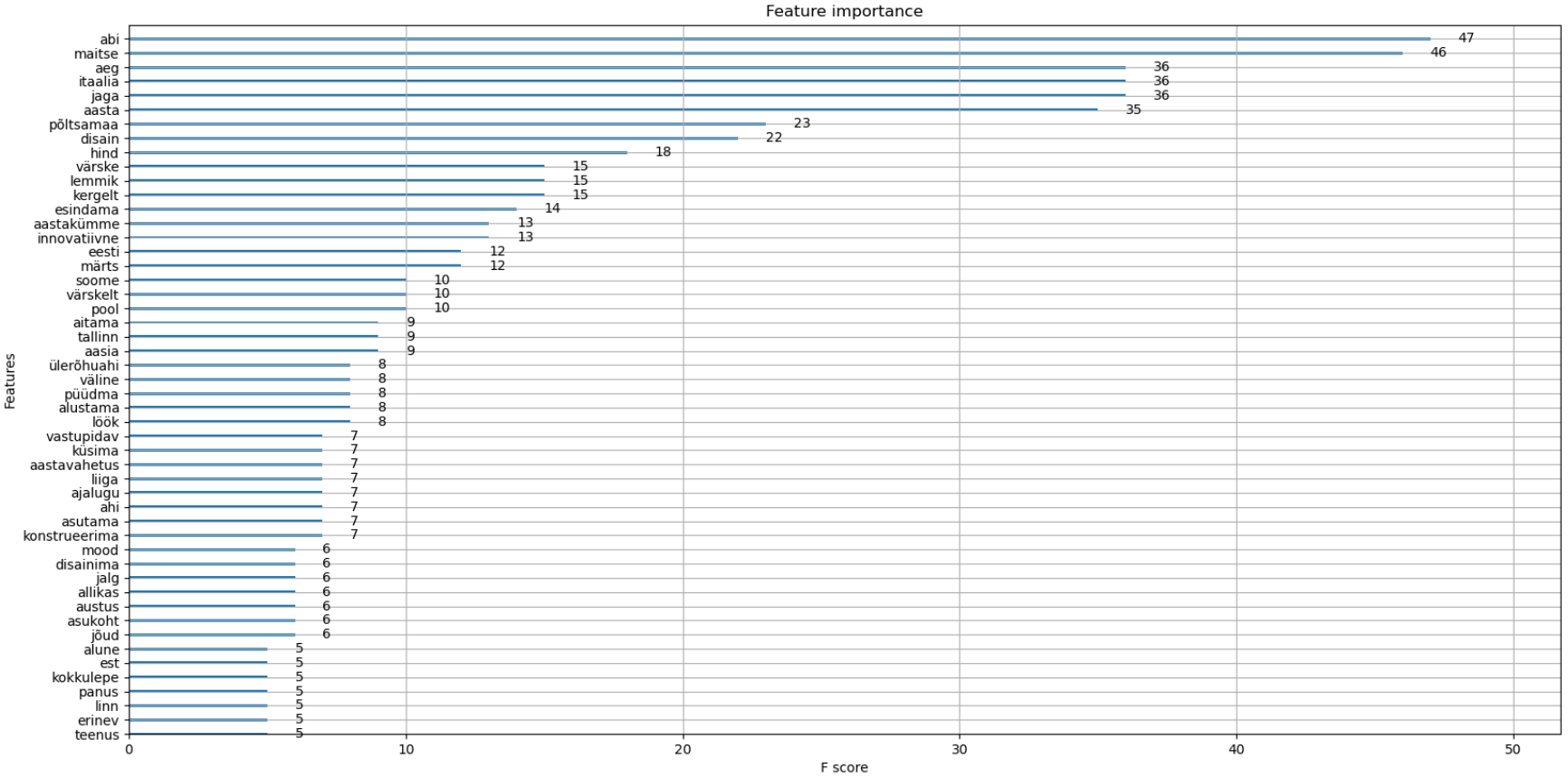


Figure 5. Feature of importance plot, 50 highest values with GX Boost model. and Random forest (n=40) (right).

Source: compiled by the author

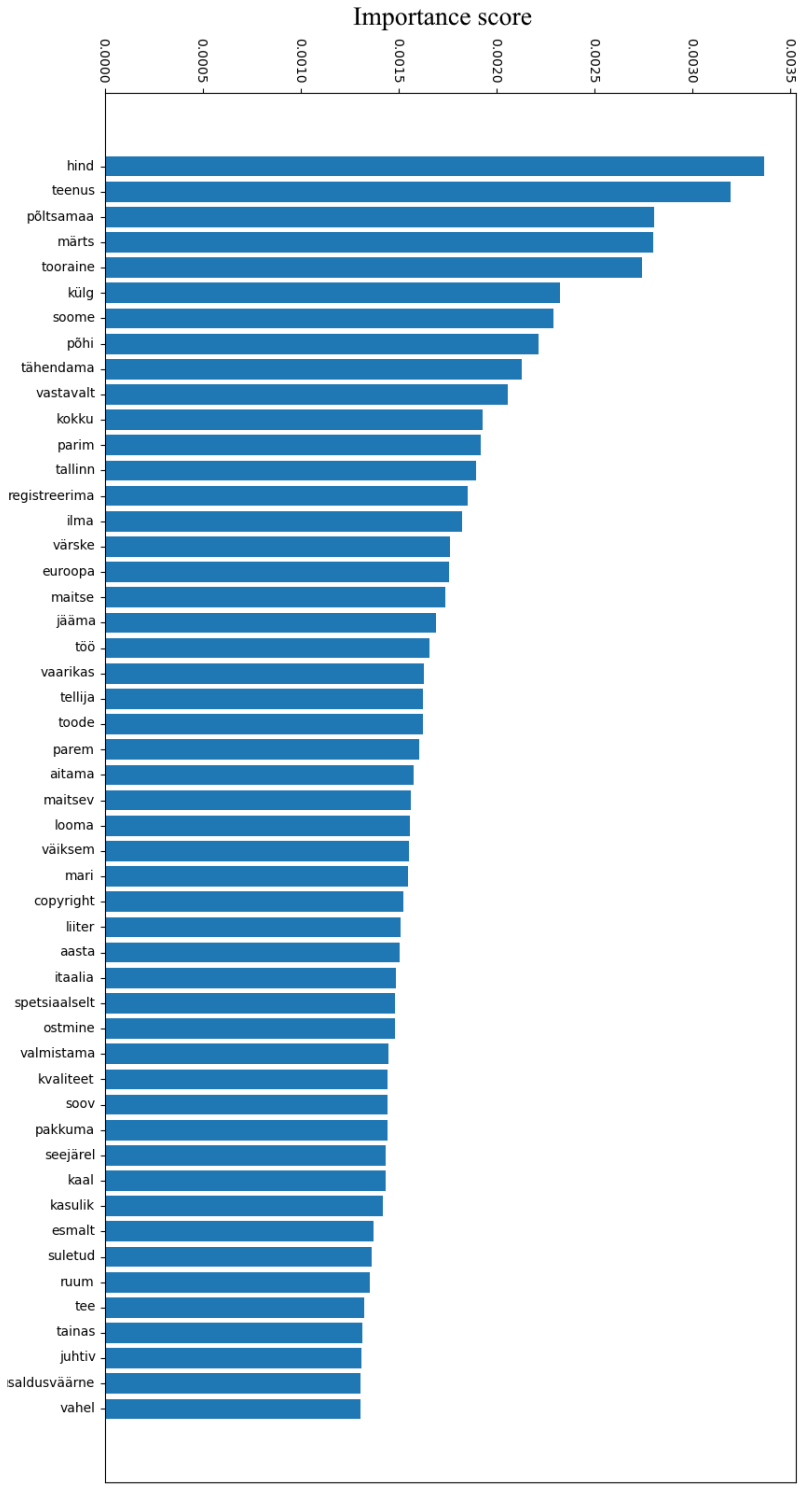


Figure 6. Feature of importance plot, 50 highest values with Random forest (n=40) model.

Source: compiled by the author

Table 8. Category of top 50 words from XG Boost and Random Forest (n=40) models. (in both models)

Category	**Exact words what were in both top 50 models**
Product characteristics (3)	hind
	värske
	maitse, maitsev
	teenus
Places (4)	Itaalia
	Soome
	Tallinn
	Põldsamaa
Time (2)	aasta
	märts

Source: compiled by the author

Tabel 9. Category of top 50 words from XG Boost and Random Forest (n=40) models. (only in one model)

Category	**GX Boost**		**Random Forest (n=40)**	
Product characteristics (22)	kergelt innovatiivne värskelt vastupidav jõud alune austus	spetsiaalselt jalg	külg põhi parim parem maitsev väiksem liiter	kvaliteet kaal kasulik suletud juhtiv usaldusväärne
Product itself (9)	ahi ülerõhuhi		ruum tee tainas mari	vaarikas tooraine toode
Marketing words (9)	disain mood disainima	lemmik esindama	ostmine registreerima	soov tellija
Costumer support (Helping) (6)	abi aitama konstrueerima	jaga küsima	aitama	
Time (3)	aastakümme aastavahetus	aeg		
Places (6)	linn est asukoht	Aasia Eesti	Euroopa	
Rights (3)	copyright kokkulepe	allikas		
Other words (18)	väline liiga erinev panus		ilma seejärel esmalt vahel kokku vastavalt püüdma	löök jääma tähendama töö looma valmistama pakkuma

Source: Compiled by the author

In table 9 all the words were divided into categories, but they were not given a separate meaning, because they were in only one model top 50 words. So, only the categories were analyzed and described. The biggest category was product characteristics with 22 words. Product characteristics and the product itself were representatives of the product innovation (combined 33 words). It proves that most probably with this method product innovation is easiest to study. Second were words that were related to marketing and customer support. It could be labeled as marketing innovation. The last category was "Other words". These were the words that may have universal meaning or words that go with other words to have a meaning, so they were drop out. It is interesting to note that one model had importance for Asia, and another had Europe. If the company exports to the whole continent it is highly possible that this company will do product innovation. Let us remember that innovation is from the definition of CIS and producing anything new is innovative.

The goal of this paper was to get a working model with the neural network. This model is not a ground truth, but it has a decent score of 0,62. So, almost 2 out of 3 predictions from it are right. Because ground truth was not discovered, the next step was to analyze different types of models and reporting their performance. With that, it was possible to get two best models with comparable AUC ROC scores. Then feature of importance was applied to the models to get the top 50 words from both models. Finally, the words with top importance were analyzed and the description for them was applied. With web text data it is possible to analyze product innovation because different models will weight the words that are associated with this innovation. Paper does not describe words about process innovation because there were fewer words than in product and marketing innovation. It is also possible to learn marketing innovation with web text but as shown above, it would be easier to learn product innovation behind it.

CONCLUSION

Over the last decade, computational power for machine learning has been decreasing which makes it more attractive for the newcomers. At the same time, it has allowed letting machine learn text as data. This is a new trend at least for social studies. In innovation studies, there has been a great use of neural networks, but not many with the website texts. Articles in online are in a structured way on structured webpages. Companies webpages are done in the "unstructured" way, which means that every webpage has its own html structure and it makes collecting data a whole lot harder because after scraping the data there is more unwanted text in, like (html code pieces). So, it takes more time to filter and prepare for the neural network. Text overall for machine learning has very high dimensionality. That is a hard part but at the same time can be a blessing. There is a possibility that text will describe a phenomenon like innovation better than traditional ways.

This paper discusses that traditional indicators may not describe innovation enough and there is a need for new kind on indicators. R&D and patent indicators are so hard to define just with numerical indicators or have too large of the time lag with them. These are called objective approaches because they try to find ground truth. But that makes them hard to use in an ever-accelerating pace of improvements. These indicators are meant to cover radical innovations, but to learn novelty we need to understand incremental changes too. Innovation studies in Europe are conducted every two years with a community innovation survey (CIS). Every country will ask their companies questions about four types of innovation: product, process, organizational, and marketing innovation. This way CIS can cover more data about the companies. Problems with that are that it takes a long time and money to make these surveys. The sample size is limited to how much data is collected with phones or e-mail. At the same time, interviewees may understand questions wrongly, because of the cultural, organizational, or economical environment. That is why collecting data on the web would be better. Because it is possible to collect data from every participant who has a webpage. In this decade for a successful company, it is

necessary to have a webpage. It is like a business card of the company where the clients go to look at their products or services.

Data collection started at December 2019. First step was to get companies data from the databases. Priority was companies' webpage but for the filtering companies background data was also needed, like how many workers company have or what is their turnover, revenue. After filtering, 2684 companies who were active and had webpages were scraped with programme called ARGUS. It gave a csv-file as an output, but it did not filter data in Estonian. Webpage text collection and filtering was done in February 2020 till end of May 2020. Natural language processes were used on the text data and labels were given CIS survey results from the year 2016 and 2018. There were 633 manufacturing companies from the both CIS-s who were labelled with product innovation 1 (innovative) or 0 (non-innovative). All data was converted into tf-idf vectors matrix. That data was validated at accuracy in the area of between 0,55-0,65 in different configurations and models. These accuracies were reported and best of configurations were tested with AUC ROC scores to see what models where better. Best model configuration were with tf-idf vector with vocabulary limit of minimum 5 words and with k-fold cross valuation set to 4. Best models were forest classification with n-estimators 40 and GX Boost models. With both models feature of importance plot was graphed and analysed. Then top 50 words from both models were compared and categorised.

Words what were in both models were about products, region, and time. Categories of the words were about products, marketing, region, time, and rights. 1/3 on all the model top 50 words from both models were associated with products what means product innovation can be learned in the machine learning and neural networks. Secondly marketing innovation would be easier to study than process or organisational innovation. Website data provided that most commonly in Estonia manufacturing companies are exporting and working closely with Finland and Italy. One of the surprises was when in the time category a month of March was as a word with the importance of the innovative companies. Because 2020 corona pandemic hit Estonia in March that can be why innovative companies were making more public reports about that month. So, online, and up to date info is can be found with the companies with the product innovation.

Neural networks and Artificial Intelligent (AI) methods with the text as data are promising. But they are very time consuming. Filtering and applying natural language processes for the text can take time and possibilities for the analyses are wide. Because this area is progressing like innovation itself, there will be always new methods to learn. But to compare to traditional methods it can still be faster and cover more total sample. Limitations to the webpages are firstly, that some company's homepages have more picture with text than text itself. They might be innovative and have a novel homepage, but for this study they were filtered out, because they did not have html text. That is will make data scraping a lot of harder because it needs then AI to understand pictures. Secondly one concern is that companies do not put sensitive info out to their homepage. Especially information about their new processes or innovation. But It can be argued that they must market they products and services and it is needed to write about every incremental change. For further research it should be interesting to conduct an innovation study for every Estonia business sector and see the product innovation in these. Secondly it is possible to analyse Facebook or other social media because lot of companies who had web pages had social media pages too.

REFERENCES

1) Archibugi, D. and Planta, M., 1996. Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), pp.451-519.
2) ARGUS: Automated Robot for Generic Universal Scraping, datawizard1337, Github, [https://github.com/datawizard1337/ARGUS] 11.06.2020.
3) Argyres, N.S. and Silverman, B.S., 2004. R&D, organization structure, and the development of corporate technological knowledge. *Strategic Management Journal*, *25*(8-9), pp.929-958.
4) Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J. and Dedene, G., 2002. Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*(1), pp.191-211.
5) Bogers, M. and Lhuillery, S., 2011. A functional perspective on learning and innovation: Investigating the organization of absorptive capacity. *Industry and innovation*, *18*(6), pp.581-610.
6) Bogliacino, F., Perani, G., Pianta, M., Supino, S., 2009. Innovation in Developing Countries. The Evidence from Innovation Surveys. Paper for the *FIRB conference. Research and Entrepreneurship in the knowledge-based economy* (p. 25).
7) Bunderson, J.S. and Sutcliffe, K.M., 2002. Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. *Academy of management journal*, *45*(5), pp.875-893.
8) Bönte, W. and Keilbach, M., 2005. Concubinage or marriage? Informal and formal cooperations for innovation. *International Journal of Industrial Organization*, *23*(3-4), pp.279-302.
9) Chen, E., Gavious, I. and Lev, B., 2017. The positive externalities of IFRS R&D capitalization: enhanced voluntary disclosure. *Review of Accounting Studies*, *22*(2), pp.677-714.

10) Choudhary, M.A. and Haider, A., 2012. Neural network models for inflation forecasting: an appraisal. *Applied Economics*, *44*(20), pp.2631-2635.
11) CIS 2018, Eurostat, Last change 05.05.2020. [https://webgate.ec.europa.eu/fpfis/wikis/display/RMSDE/CIS+2018]. 03.08.2020.
12) Clausen, T., Fagerberg, J. and Gulbrandsen, M., 2012. Mobilizing for change: A study of research units in emerging scientific fields. *Research policy*, *41*(7), pp.1249-1261.
13) Clem, A., Cowan, A.R. and Jeffrey, C., 2004. Market reaction to proposed changes in accounting for purchased research and development in r&d-intensive industries. *Journal of Accounting, Auditing & Finance*, *19*(4), pp.405-428.
14) Collobert, R. and Weston, J., 2008, July. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167).
15) Community Innovation Survey (CIS), Description of dataset, Eurostat [https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey] 08.06.2020.
16) Description of dataset, Community Innovation Survey (CIS), Eurostat, [https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey] 14.06.2020.
17) Ettevõtete innovatsiooniuuring. aastad 2016-2018, Statisikaamet küsimustiku kood: 12932019. [https://www.stat.ee/1426704]. 06.06.2020
18) Eurostat, Community Innovation Survey (CIS) microdata, Note to the CIS researcher visiting Eurostat's SAFE Centre, 2013. [https://ec.europa.eu/eurostat/documents/203647/203701/Note-CIS-researcher-Eurostat-SAFE-Centre.pdf/2529ad58-aee3-4cb9-9da2-094563ad0fae]. 15.06.2020.
19) Fagerberg, J., 2013. TIK WORKING PAPERS on Innovation Studies, pp. 1–45..
20) Fisher, J., Craig, A. and 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.

21) Gault, F. ed., 2013. *Handbook of innovation indicators and measurement*. Edward Elgar Publishing, p. 512.
22) Geuna, A., Kataishi, R., Toselli, M., Guzmán, E., Lawson, C., Fernandez-Zubieta, A. and Barros, B., 2015. SiSOB data extraction and codification: A tool to analyze scientific careers. *Research Policy*, *44*(9), pp.1645-1658.
23) Gentzkow, M., Kelly, B. and Taddy, M., 2019. Text as data. *Journal of Economic Literature*, *57*(3), pp.535-574.
24) Godin, B., 2002. The rise of innovation surveys: Measuring a fuzzy concept. *Canadian Science and innovation indicators consortium, project on the history and sociology of S&T statistics, Paper, 16,* p. 26.
25) Griffith, R., Huergo, E., Mairesse, J. and Peters, B., 2006. Innovation and productivity across four European countries. *Oxford review of economic policy*, *22*(4), pp.483-498.
26) Gök, A., Waterworth, A. and Shapira, P., 2015. Use of web mining in studying innovation. *Scientometrics*, *102*(1), pp.653-671.
27) Henderson, R. and Cockburn, I., 1994. Scale, scope and spillovers: the determinants of research productivity in ethical drug discovery. . *The Rand journal of economics* , *27,* pp. 32–59.
28) History and License, Python documentation, Python. [https://docs.python.org/3/license.html] 10.06.2020.
29) Kim, Y., 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
30) Kinne, J. and Axenbeck, J., 2018. Web mining of firm websites: A framework for web scraping and a pilot study for Germany. *ZEW-Centre for European Economic Research Discussion Paper*, (18-033) p. 47.
31) Kinne, J. and Lenz, D., 2019. Predicting innovative firms using web mining and deep learning. *ZEW-Centre for European Economic Research Discussion Paper*, (19-01), pp. 1–10.
32) Klette, T.J., 1996. R&D, scope economies, and plant performance. *The RAND Journal of Economics*, pp.502-522.
33) Kleinknecht, A. and Reijnen, J.O., 1991. More evidence on the undercounting of small firm R&D. *Research Policy*, *20*(6), pp.579-587.

34) Klevorick, A.K., Levin, R.C., Nelson, R.R. and Winter, S.G., 1995. On the sources and significance of interindustry differences in technological opportunities. *Research policy*, *24*(2), pp.185-205 referred through Laursen, K. and Salter, A., 2006. Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strategic management journal*, *27*(2), pp.131-150.
35) Laursen, K. and Salter, A., 2006. Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strategic management journal*, *27*(2), pp.131-150.
36) Levenberg, A., Pulman, S., Moilanen, K., Simpson, E. and Roberts, S., 2014. Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*, *3*(2), pp.109-115.
37) Lhuillery, S., Raffo, J. and Hamdan-Livramento, I., 2016. *Measuring creativity: Learning from innovation measurement* (Vol. 31). WIPO, pp .1–24.
38) Mairesse, J. and Mohnen, P., 2010. Using innovation surveys for econometric analysis. *CIRANO-Scientific Publication*, (2010s-15), pp. 1–40..
39) Mohr, L.B., 1969. Determinants of innovation in organizations. *American political science review*, *63*(1), pp.111-126.
40) Nagaoka, S., Motohashi, K. and Goto, A., 2010. Patent statistics as an innovation indicator. In *Handbook of the Economics of Innovation* (Vol. 2, pp. 1083-1127). North-Holland.
41) MS Excel support [https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3] 05.03.2020.
42) OECD Publishing, 2018. *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation*. Organisation for Economic Co-operation and Development OECD.
43) OECD Publishing, 2015. *The Future of Productivity.*
44) OECD Publishing, 2002. *The Measurement of Scientific Technical Activities. Frascati Manual.*
45) OECD Publishing, 2005. *Oslo Manual, Guidelines For Collecting And Interpreting Innovation Data*.

46) Polder, M., Van Leeuwen, G., Mohnen, P. and Raymond, W., 2009. Productivity effects of innovation modes. *Statistics Netherlands*, p. 24.

47) Rao, A. and Spasojevic, N., 2016. Actionable and political text classification using word embeddings and lstm. *arXiv preprint arXiv:1607.02501*.

48) Raymond, W., Mohnen, P. A., Palm, F. Loeff, van der S. S., 2006. An empirically-based taxonomy of Dutch manufacturing: innovation policy implications. *Leibniz Institute for Economic Research at the University of Munich*, 1230, p.38.

49) Sauermann, H. and Roach, M., 2013. Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, *42*(1), pp.273-286.

50) Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural networks*, *61*, pp.85-117.

51) Schumpeter, J.A., 1954. History of economic analysis. Psychology Press, p. 1283.

52) Smith, K., 2009. Measuring Innovation. *The Oxford Handbook of Innovation*, pp. 148–173.

53) Teatmik.ee database, 2020. [https://www.teatmik.ee/et/advancedsearch]. 10.06.2020.

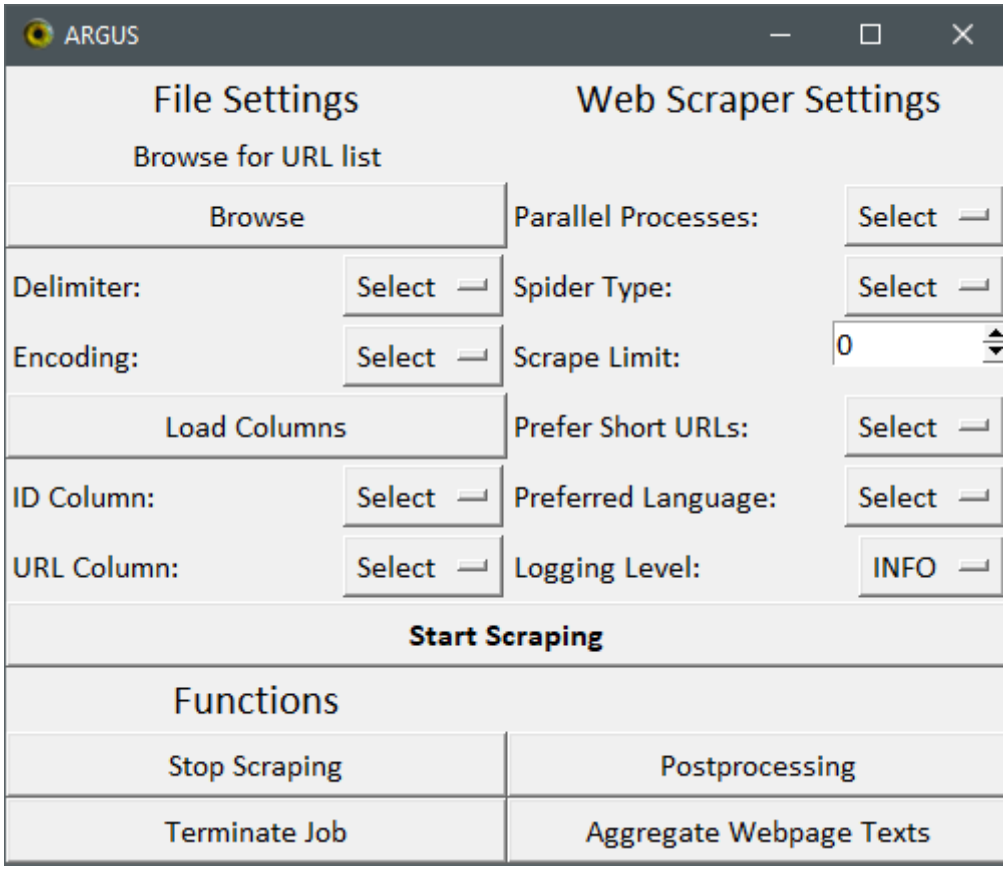
54) Toll, M., 2017. DIY Lithium Battaries, How to Build Your Own Battery Packs, p. 221.

55) 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.

56) Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., 2016, Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480-1489).

APPENDICES

Appendix 1. ARGUS layout.



Sources: (ARGUS: Automated...: 2020)

Appendix 2. Manufacturing companies' webpage coverage by the Estonia counties in 2020 January.

County	Nr of companies	Companies who have website	%	Companies who have 5 or more emplyees and a website	%
Valgamaa	241	82	34%	33	14%
Võrumaa	371	130	35%	56	15%
Hiiumaa	140	40	29%	13	9%
Jõgevamaa	297	106	36%	46	15%
Lääne-Virumaa	630	233	37%	98	16%
Põlvamaa	296	103	35%	37	13%
Saaremaa	508	192	38%	78	15%
Tartumaa	2460	770	31%	298	12%
Pärnumaa	812	292	36%	118	15%
Ida-Virumaa	1010	290	29%	146	14%
Viljandimaa	448	137	31%	63	14%
Raplamaa	502	172	34%	67	13%
Läänemaa	295	91	31%	37	13%
Järvamaa	329	120	36%	58	18%
Harjumaa	9308	3722	40%	1536	17%
Sum	**17647**	**6480**	**37%**	**2684**	**15%**

Source: compiled by the author

Appendix 3. Summary of economically active manufacturing companies and companies with more than 1 employee in 2019.

County	Nr of companies who have more than 1€ revenue	Nr of companies who have at least 1 emplyee	Nr of active companies
Valgamaa	111	109	88
Võrumaa	218	223	184
Hiiumaa	72	67	54
Jõgevamaa	192	199	172
Lääne-Virumaa	384	395	329
Põlvamaa	170	164	135
Saaremaa	277	269	221
Tartumaa	1308	1375	1178
Pärnumaa	412	419	341
Ida-Virumaa	591	660	510
Viljandimaa	256	254	204
Raplamaa	317	312	264
Läänemaa	175	173	151
Järvamaa	215	212	186
Harjumaa	5518	5321	4549
Sum	**10216**	**10152**	**8566**

Sources: compiled by the author

Non-exclusive licence to reproduce thesis and make thesis public

I, Sander Sõna (08.12.1991),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

„PREDICTING INNOVATING COMPANIES IN ESTONIA BY ANALYSING MANUFACTURE COMPANIES WEBSITE DATA“

(title of thesis)

Supervised by Supervisors: Jaan Masso, Rajesh Sharma, Priit Vahter,

(supervisor’s name)

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 11.08.2020