Tampereen yliopisto

Aki Haanpää

APPLYING NATURAL LANGUAGE PROCESSING IN TEXT BASED SUPPLIER DISCOVERY

Faculty of Engineering and Natural Sciences
Master's Thesis
Jussi Heikkilä
Juho Kanniainen
November 2019

TIIVISTELMÄ

Aki Haanpää: Luonnollisen kielen prosessoinnin soveltaminen tekstipohjaisessa toimittajan löytämisessä
Diplomityö
Tampereen Yliopisto
Tuotantotalouden DI-tutkinto-ohjelma
Marraskuu 2019

Siinä missä toimittajavalintaa on tutkittu laajalti, ei toimittajan tunnistamiseen ja löytämiseen liittyviä prosesseja niinkään. Toimittajan tunnistaminen ja ylipäätänsä löytäminen on osa toimittajavalintaan liittyvää prosessia. Toimittajan tunnistamiseen liittyvät prosessit ovat vielä melko alkuvaiheessa ja näin ollen olisivat kehitettävissä. Kuitenkin, yritysten keskuudessa on kasvanut kiinnostus kehittää toimittajan tunnistamiseen liittyviä prosesseja eri toimialoilla.

Yritykset keräävät valtavia määriä tietoa näiden ostoihin ja hankintaan liittyvistä prosesseista erilaisiin tietojärjestelmiin. Pääosin yritysten säilömä data sisältää tietoja ostotilauksen kuvauksesta, eli siitä mitä ja mistä jotakin on ostettu ja kuka on ollut kyseinen toimittaja. Tämä hankintojen sisältämä tieto on ihmiselle ymmärrettävässä tekstimuodossa. Tämän tekstidatan hyödyntämisessä saattaa piillä valtava liiketoiminnallinen potentiaali tai mahdollisuus uuden tiedon luontiin. Tässä tutkimuksessa mukana ollut Sievo on yksi näistä yrityksistä, joka on pyrkinyt ajamaan tämän datan hyödyntämistä liiketoiminnassa.

Tämän diplomityön alkuvaiheessa koneoppiminen nähtiin järkevänä vaihtoehtona datalouhinnan työkaluksi. Yksi koneoppimisen kehittyneimmistä tekniikoista on luonnollisen kielen prosessointi menetelmä. Tämän pohjalta, yhdessä Sievon kanssa alettiin tutkia, voidaanko tekstipohjaista ostotilauskuvausta käyttää luonnollisen kielen prosessoinnin lähteenä siten, että sillä pystyttäisiin tuottamaan arvoa toimittajanlöytämisprosessiin. Ajatuksena on, että koneoppimista hyödynnetään tekstidatan käsittelyn automatisoinnissa. Aikaisen vaiheen testissä sovellettiin kahta koneoppimistekniikkaa, fastText-algoritmia sanavektorien muodostamiseen ja HDBScania vektoreiden kulsterointiin. Näin tarkoituksena oli yhtäläisyyksiä ostotilausten ja näitä vastaavien toimittajien välillä. Näiden yhtäläisyyksien perusteella pystyttäisiin tunnistamaan uusia mahdollisia toimittajia yrityksille.

Tutkimuksen perusteella voidaan todeta, että fastText ja HDBScan onnistuivat tuottamaan järkeviä tuloksia, sekä tunnistamaan uusia toimittajia datasta. Lopputuloksena tutkimus tuotti klusteroimalla ryhmiteltyjä joukkoja samankaltaisia toimittajia ostojen tapahtumakuvauksien perusteella. Havaintoja arvioitiin yhdessä Sievon asiakkaan kanssa, joka on samalta toimialalta kuin tutkimuksessa hyödynnetty data.

Kuitenkin, tutkimuksen valossa voidaan todeta, että vaikka algoritmit onnistuivatkin tuottamaan ymmärrettäviä tuloksia ja listaamaan uusia toimittajia, ostotilausten sisältämät tekstipohjaiset kuvaukset eivät riitä tuottamaan riittävän rikkaita löydöksiä. Listaus uusista toimittajista ei riitä antamaan tarpeeksi tietoa toimittajavalinnan päätöksenteon tueksi. Toisaalta asiakkaan mukaan tällaisen koneoppimiseen pohjautuvan menetelmän hyödyntäminen voisi kehittyä tulevaisuudessa prosesseihin arvoa tuottavaksi. Tutkimukseen nojaten, on syytä toteuttaa jatkotutkimuksia erilaisten koneoppimismenetelmien vaihtoehtojen selvittämiseksi. On myös syytä tutkia, kuinka tekstikuvauksien rikastaminen vaikuttaisi tutkimustuloksiin. Tällä hetkellä tekstitiedon vajavuus vaikuttaa tarkkuuteen.

Avainsanat: Toimittajan löytäminen, toimittajan tunnistaminen, toimittajavalinta, luonnollisen kielen prosessointi, koneoppiminen, tekstinlouhinta, fastText, HDBScan

ABSTRACT

Aki Haanpää: Applying natural language processing in text based supplier discovery
Master's Thesis
Tampere University
Degree Programme in Industrial Engineering and Management, MSc (Tech)
November 2019

Supplier selection has been widely studied research area, whereas supplier discovery in turn is not. Being part of the supplier selection process, supplier discovery aims to identify new prospective suppliers from the mass. Increasingly, from the business point of view, supplier discovery process lives still its infancy and can be developed further. Seemingly, an incentive to improve the process efficiency has emerged among the companies in different industries.

Companies collect data and information about their purchases and other procurement operations, hence a massive amount of data exists in their procurement ERP-systems. The data generally includes e.g. purchase order descriptions, what has been bought and sourced, and from which supplier. This information the purchase orders contain is in textual, human understandable format. From a business point of view using this data may have business leveraging potential and create new knowledge, as is some companies might be willing to utilize the data. Sievo among the others, has been driving this incentive to utilize the data.

In this thesis, it was initially seen that machine learning techniques could likely offer a reasonable option for successful data mining tool. One of the sophisticated techniques is natural language programming, NLP. On these basis, together with case company Sievo, were launched the thesis research project as target to examine, if textual purchase order descriptions could be utilized with NLP techniques to deliver extensive value to supplier discovery process by automating the purchase order description mining. Two machine learning techniques were implemented to conduct an early-stage test, fastText used for creating word vectorizations and HDBScan for clustering the word vectors. The idea was then to find possible similarities between different transactions, that would then relate further to similar kind of suppliers. By extracting the similarities between different companies' purchase order would enable to identify group of new suppliers.

The research shows that implemented methods, fastText and HDBScan, were able to conduct meaningful results, and to identify new suppliers from the text data. During the implementation the textual data was fed through the fastText algorithm and eventually the HDBScan clustered the similarities. In the end, the research generated clusters with groups of similar kind of suppliers by transaction descriptions. Observations were evaluated together with one Sievo client, representing the same industry field as the sample data was taken from, telecom industry.

Apparently, the research implications concluded that even though the implemented algorithms were able to conduct comprehensible results and list of new suppliers identified, the text data the purchase orders contained was insufficient enable delivering critical information to the decision-makers in charge of the supplier selection. However, the customer agreed that these kind of text mining might have potential to evolve into an applicable product eventually, as it now already offers extensive information for the decision-makers and hence is able to create some value. However, in the light of research evidence it is suggested to perform further research to test and benchmark different alternatives of creating the word vectors and for clustering them as well. Additionally, the transaction descriptions should be enriched, as now not all the rows were descriptive enough for algorithms to be able to filter out exact similarities.

Keywords: supplier discovery, supplier identification, supplier selection, natural language processing, machine learning, text mining, fastText, HDBScan

PREFACE

It is often said that nothing good comes easy, nor comes worth having. Now, already a year ago since I started this seemingly massive and time-consuming research project is drawing to a close. A lot have been studied, a lot have been learned. Even more experienced I am now than over a year ago. Nevertheless, it sometimes felt like the Pareto principle, 80/20 rule, as the majority of the contribution to this project accumulated during the 20% of the latest time period still with an emphasis of 80% regarding the content and magnitude, I'm confident to say it all turned out well.

I would like to address my humble gratitude to the most helpful and supporting people on this journey. Firstly, I'd like to thank my Professors Jussi Heikkilä and Juho Kanniainen for giving explicit academic guidance with the thesis, already during the university years, also for bearing my almost infinite project and all those fluctuating moments of frustration and uncertainty. Probably not the ideal example of executing a Master's Thesis project. Still, your support has been essential. Secondly, my sincerest gratitude belongs to my beloved family, who has been supporting me every single moment during my thesis writing, and beyond.

Not should be left out mentioning my dearest friends: fellow study buddies, colleagues at work, and people outside the academic and professional context. Distinctively desire to give recognition to Jeris, Roope and Minttu for sharing the severity embraced and giving a helping hand, whenever it was needed. Appreciation should be expressed also to Sanna (the best colleague one can have) and Jesse Saari who both unselfishly stood by my side no matter what happened with my thesis or in my life. Additionally want to thank the rest of my colleagues at Sievo (especially Sammeli and Matan for helping me with the actual research), Blebeijikerho, SilverPlanet, EK, T-hill, Slush Partnerships Team aka Wolfpack, Tomi and all those other individuals who have contributed to this thesis.

In Helsinki, Finland on 25 November 2019

Aki Haanpää

SISÄLLYSLUETTELO

LIST OF SYMBOLS AND ABBREVIATIONS

AHP	Analytical Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANP	Analytic Network Analysis
B2B	Business-to-Business
CBOW	Continuous Bag-of-Words
CSV	Comma-Separated Value
DBScan	Density-Based Spatial Clustering of Application with Noise
DEA	Data Envelopment Analysis
DL	Deep Learning
DSM	Distributional Semantic Models
ERP	Enterprise Resource Planning
FMCG	Fast-Moving Consumer Goods
GA	Genetic Algorithm
HDBScan	Hierarchical Density-Based Spatial Clustering of Applications with Noise
KNN	K-Nearest Neighbor
MCDM	Multi-Criteria Decision-Making
ML	Machine Learning
NLP	Natural Language Processing
PCA	Principal Component Analysis
POLineDesc	Purchase Order Line Description
SQL	Standard Querying Language
SVD	Singular Value Decomposition
Telco	Telecommunication
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
t-SNE	t-Distributional Stochastic Neighbor

A	vector A
A_i	component of vector A
B	vector B
B_i	component of vector B
θ	angle

.

1. INTRODUCTION

This research was conducted in a close collaboration with a Finnish procurement analytics software company, Sievo, or later *case company*. As the initial preliminary interest emerged from their business and product development needs, and thus together the study objectives were then elaborated and determined resulting to quite a practical master thesis work. There is an immense amount of data stored in Sievo data warehouses consisting of customers' procurement data just waiting to be exploited and processed for further business intentions. For some time now Sievo has been planning to examine if that data and strategic supplier selection could be interconnected in some way. For Sievo the main goal was to identify if any potential business opportunities exist within constantly evolving field of data analytics and applied machine learning in context of procurement. From the academic point of view, the core research incentives were to explore whether natural language processing techniques could leverage the supplier discovery process and be utilized for value creation in regarding. Here will be introduced the managerial motivations and academic canvas for conducting this master thesis.

1.1 Research motivation, objectives and scope

Whereas supplier selection has been acknowledged as a critical part of supply chain management, in turn not quite widely studied (Wetzstein *et al.* 2016) area is supplier discovery. Supplier discovery, or supplier identification as both the terms appear interchangeably in literature (Kang *et al.* 2011; Lee *et al.* 2011; Wetzstein *et al.* 2016), describes the process of initial screening of a prospect supplier before proceeding in to particular evaluation phase. It has been recognized that this discovery process can be very challenging and encompass significant amount of resources. In general, the exhausting and time-consuming factors are highly related to uncertainty or lack of information (Cheraghi *et al.* 2004; Sarkis & Talluri 2002). In order to efficiently tackle these problems, a systematic solution with up-to-date information is required.

During the past few years sophisticated applications of data analytics have become more common in supply chain management and procurement context. Furthermore, organizations store an increasingly amount of data in their IT systems, hence a seemingly large database exists. This has increased the interest of efficient utilization of the stored data in procurement business processes. As a tremendous amount of procurement data flows through different information systems among various companies, obviously consisting information about their suppliers and purchased transactions, the main incentive here was to examine whether that data could be utilized explicitly in supplier discovery, support in the decision-making process or other closely related tasks. Furthermore, the spend data that flows through enterprise supply chain management systems embodies lots of textual data about the items and goods purchased. Expectedly various suppliers are able to deliver same products and materials, and hence mapping probable hidden synergy benefits would be interesting.

When speaking of data analytics nowadays, artificial intelligence (AI) and machine learning (ML) are often appearing concepts with multiple of ambiguous variant definitions (Bokka *et al.* 2019; Beysolow II 2018; Cavalcante *et al.* 2019; Chopra *et al.* 2016; Kao *et al.* 2007). These two entities involve a wide range of different popularly known and unknown methods and techniques which have been commonly implemented already in many field of businesses and sciences. Particularly in commercial industry like entertainment service systems and ecommerce platforms (e.g. Netflix, Facebook or Zalando) extensive exploiting of the available user data has been identified as a very considerable key factor in recommending relevant new products and services to the customers in question. This has partly lead also B2B sector to implement these originally for commercial use developed methods, and to evolve towards more data-driven business model.

Apparently some of these kind of implementations have already been utilized at Sievo. Furthermore, from Sievo side data scientists, data engineers and experts were received plenty of recommendations and encouragement to investigate the possibilities to apply for example AI methods or deep learning, a collection of machine learning algorithms used for extracting top level features from an unprocessed dataset, in this research. Distinctively interest towards exploitation of textual procurement data and purchase order lines.

As such, this thesis aims to examine if procurement data, particularly spend data gathered from specific Sievo customers' could be used as source for text similarity based supplier discovery. Ergo, the main goal of the research was to study if new suppliers could be identified by comparing text similarities between different purchase order descriptions, and moreover whether the textual spend data is adequate to generate sufficient discoveries. Enabling to chart purchase order line similarities Sievo data experts advised to apply natural language processing techniques. Thus, this thesis scrutinizes as well the feasibility of using fastText, the implemented NLP method in question, with purchase order text from the business value perspective. The performance of fastText will be evaluated through business value creation viewpoint, thus one Sievo customer inputs feedback regarding the derived results. Therefore, two high-level research questions were derived and formulated as following:

1. *Can purchase order descriptions be used as source for natural language processing based supplier discovery?*
2. *How natural language processing based supplier discovery performs in bringing value in supplier selection process?*

Now, these main research questions breed a couple of considerations needed to be addressed and sharpen while answering to the key problems. Firstly, defining a successful use case of text based supplier discovery is necessary. This will be in fact evaluated together with Sievo customer by interviewing and collecting feedback. The customer operates in the same industry segment than where the implemented data was gathered from, and is thus able to deliver insights and knowledge in this regard, and analyze the results of text based discoveries

Similarly, considering the second research question, one of the main incentives in this research is to examine how particularly the text based supplier discovery would perform in creating new supplier suggestions from supplier selection value creation point of view. Can be rephrased as if using text based supplier discovery succeeds to identify new suppliers initially, should be evaluated whether the derived information is useful or is able to bring additional value to the supplier selection process. Assuming that if new suppliers could be identified, it is reasonable to assess the degree of informativeness and implicativeness of the derived results. Whether that new information would support in supplier selection process, and moreover estimate the perceived benefits and business impact from the managerial decision-making point of view. Hence, it is required to cover what underlying factors play a role in value creation of supplier selection from customer's supplier management perspective.

Nonetheless, the main research questions are formulated as such additionally necessary is needed to be covered followings: how value is created in supplier selection process? What factors affects the value creation in supplier selection? What factors affects the feasibility of using purchase order descriptions for text based supplier discovery? What alternatives are there to the algorithms used in this research? Striving for a solid outcome in the research also some limitations are determined concerning the used methods and the focus of the study:

- In this research only Sievo recommended NLP techniques are used, ergo the used techniques fastText and HDBScan were selected in accordance with Sievo data experts reference
- This research does not offer any further elaboration about the technical structures behind the algorithms not from mathematical or software engineering point of view
- The used sample data is determined by Sievo data experts, thus the given data is distinctively from telecom (or *telco*) industry customers gathered during the fiscal year 2018
- Only indirect type of spend data will be used

- In this research is used extensively one Sievo customer interview feedback when evaluating and answering the research question two
- In this research the chosen programming language was Python in the light of prior research and documentation (Chopra *et al.* 2016; Hardeniya *et al.* 2016; Beysolow 2018; Goyal *et al.* 2018; Bokka *et al.* 2019)

Hence, this thesis will not cover any comparison between implemented used techniques or their alternative options. However, this will be discussed at the end though, but no further analysis about existing method range will not be handled. As already mentioned above the chosen algorithmic methodology was largely impacted by Sievo's internal data science team and the selections made were preferred by the experts in the field, no detailed examination of available machine learning algorithms for this use case will be discussed within the scope of study. The thesis only focuses on applying particularly fastText and HDBScan techniques, which were chosen in accordance to the emphasis of data scientists at Sievo, ergo the algorithmic implementation was partly driven by case company's internal product development team professionals. The structure can be divided into sections like presented in Figure 1 below.

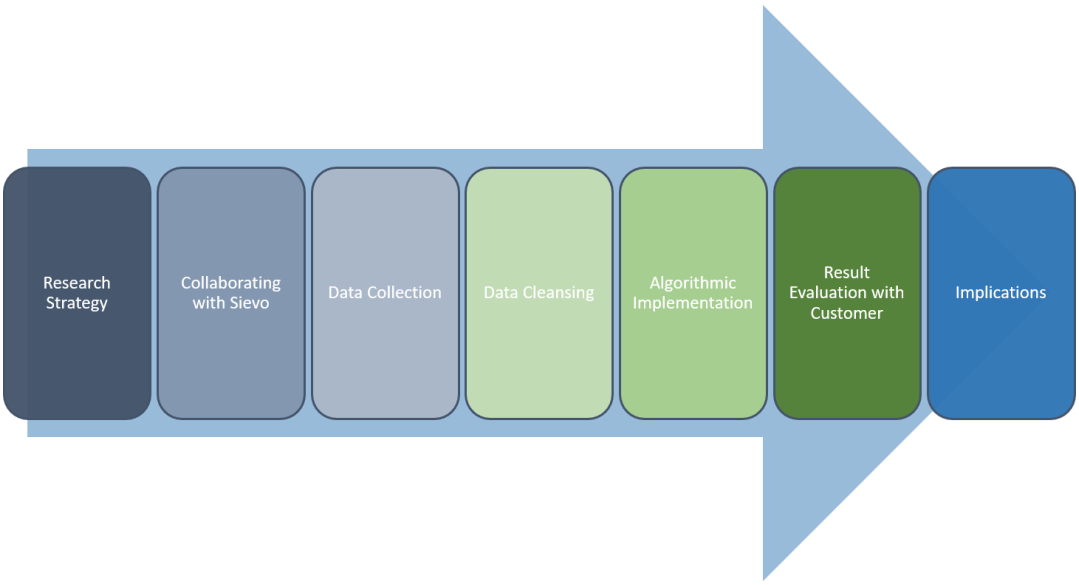

Figure 1. *Different stages during the research project*

The structure first starts with a research strategy, where initial background research about the related work in the field is carried out. Furthermore is studied some literature regarding supplier discovery and applied use cases of natural language processing in concerning context. Partly at the same time collaborating with Sievo is already started to be able to better align the business incentives with the thesis goals. After that the actual practical parts take places, where the data first gathered, used sample data cleansed and subsequently are fastText and HDBScan clustering algorithms implemented. As the data is fed through the models, next up is analysis of generated results accompanied by customer interview. The derived results will then be reflected to prior research and existing literature, as well as the available customer input, these happens in the "Implication" stage. The research implications and discussion are based on the findings accumulated during the whole process.

1.2 Case company: Sievo

Sievo is a Finnish software company operating particularly in procurement analytics field. The company has over 180 employees, mainly in Helsinki headquarters, Finland, but also in their subsidiary office in Chicago (Sievo 2019). Sievo was founded in 2003 and has current customers from every continent expect Oceania. Sievo's turnover in fiscal year 2018 was over 10M€ and

customers are from multiple different industries like telco, manufacturing and FMCG. Customer references include companies such as Carlsberg, Levi's, Deutsche Telecom, Schindler Fortum and Fiskars. Operating with international customer base in procurement analytics means that Sievo develops and delivers a title software which aims to support its global companies to optimize their procurement operations, activities and related finance. First the stored spend data in companies' ERPs are transferred to Sievo's databases, where from Sievo gains the accessibility to the raw data and which is then ready to be analyzed and processed. Sievo software lets company's procurement official or purchasers to keep track of company's spend, spend categories and possible realized savings as well as plan better campaign budgets or long term spend allocation, not to mention better visibility to supplier contract compliance and contract management. Sievo's solution enables the client companies to access perceivable structured, classified and visualized historical spend, payment terms and realized market prices through the spend data (and available extensive data) they have delivered to Sievo.

2. BACKGROUND

This chapter delivers the necessary background to be able to understand the research fundamentals. Here will be introduced the very basics of supplier discovery, its related entries, as well as what natural language processing is, and what kind of techniques and concepts are entwined. First is wrapped the conceptions that supplier discovery embodies, following the crucial phase of supplier selection, and finally natural language processing is elaborated.

2.1 Supplier discovery

Supplier discovery is a strategic supplier management process which precedes the actual supplier selection. For example, Lee *et al.* (2011) have defined supplier discovery to be collection of activities that enables identifying capable companies to deliver the desired service or goods. Ergo, when a company is facing a need to purchase or procure something e.g. item, raw material, products or services, it needs to determine where from to acquire. The need for suitable selection of supplying party requires the initial finding of the given supplier. Say that if a company is buying something usual on its daily basis, there might be already an existing short list of potential suppliers. What if regarding company wishes to buy entirely new items or products or what if it wants to put out to tender the considered procurement process. Before a company is able derive even a long list of possible suppliers, it must first identify the suppliers capable to deliver and supply the wanted service or item, ultimately. This is where supplier discovery steps in. It's highly necessary that in supply discovery process is also taken into consideration operational capabilities of screened supplier, although the technological requirements play more essential role (Ameri & McArthur 2014). Kang *et al.* (2011) divided supplier discovery into two phases:

1. Collect function
2. Search function

Subdividing supplier discovery concept in these two steps means basically, that Kang *et al.* (2011) identified that first is needed to define the basis which the actual supplier search will build upon. Likewise, during the collect function company gathers and collects data and information about the capabilities needed from the supplier. Even other stakeholders like customers or partners can be involved in this stage, delivering input to the decision-making process. After the requirements are determined comes the search part, where the suppliers best meeting the set requirements are identified. Hence, supplier selection process will supervene supplier discovery phase. This will be introduced more detailed in chapter 2.2 Supplier selection process.

In recent literature is recognized that unlike supply chain operations are widely studied area, supplier discovery instead still needs further development. Reflecting to the history, through the ages companies have purchased various items and materials from other companies. Previously, the act of finding new prominent suppliers has been done by using e.g. Yellow Pages in the phone dictionary, online search, or simply word of mouth. Not always were suppliers easy to find and sort out the most suitable for the considered business purposes. Companies already back then have struggled with the challenge regarding the search of the most valuable and perfect match supplier. There have been sometimes even cases, where a potential suppliers have been hand-picked from other industry domain. (Lee *et al.* 2011)

It is not always only the searching process that creates friction in finding the suitable supplier, additionally the level of knowledge the person responsible for tendering the suppliers has crucial part. It was discovered in a study by Mesmer & Olewnik (2018) that lacking the deep understanding of some given manufacturing process also might have an impact on the process fluency. Implying that if the process of finding prominent supplier could be automated, that'll definitely create business value and leverage the supply chain operation in question.

Prior literature supports the claim that, even though supplier discovery itself is not intensively unraveled study focus, multiple research have tried finding an automated or streamlined strategy

to sort out favorable suppliers from the whole population. For example Lee *et al.* (2011) tried classifying and filtering suppliers from supplier registry base in the respect of supplier category type, though not entirely solid solution. Also semantic rule modelling was tried in one research by Ameri & McArthur (2014), in a search method based on Manufacturing Service Description Language or MSDL. Basically MSDL was an ontology on what2 the semantic search relied. Semantic search means that the searching algorithm aims to find results by exploiting the meanings the collection of text or language embed.

However, in order to understand the role of supplier discovery from more comprehensive perspective, for instance from a company procurement officer's point of view, one is required to observe also the actual supplier selection and its decision-making process. As the supplier selection quite sets the criteria for how companies do organize the supplier discovery process. The dynamics of supplier selection process generates a couple of underlying questions: Whether new suppliers are needed to discover? What requirements are set to newly found suppliers? How the suppliers are identified as prospect supplier companies in the very beginning. Following chapter handles the entirety of actual supplier selection.

2.2 Supplier selection process

Supplier selection can be considered as an entry point of the whole supply chain process. A supplier that would be capable to deliver the purchased goods, items or materials, so that the company is able to manufacture and produce the end product finally to the customer. Many risks are related to the activity of supplier selection such like delivery reliability, delivered quality, timing, service level, supplier contract compliance, just to mention couple (Shemshadi *et al.* 2011). Within operation management, decision sciences and production economics supplier selection entity has been extensively studied (Chai & Ngai 2019). Supplier selection creates a crucial part of strategic decision-making in strategic supply chain management. In procurement, the business activity of purchasing and buying goods and services, choosing the right supplier for the different particular purposes is highly significant and can leverage organization's business performance and productivity. The prior literature (Vokurka *et al.* 1996; Tsai *et al.* 2010; Shemshadi *et al.* 2011; Ye *et al.* 2014; Yu & Wong 2014; Wetzstein *et al.* 2016; Cavalcante *et al.* 2019; Chai & Ngai 2019; Luan *et al.* 2019) supports the claim that supplier selection is one of the key activities of strategic sourcing which embodies a wide range of operations like procurement, and supplier relationship management, furthermore seen as a source of competence and building capabilities in fierce competition environment. Moreover, the selection of supplier may affect the company finance and other key operations (Vokurka *et al.* 1996). One way to approach supplier selection was introduced by Sonmez (2006). The steps are shown in Figure 2.

Figure 2. *Steps for supplier selection (Sonmez, 2006)*

Sonmez proposed that supplier selection is initiated by identifying the basic need for a new supplier, which is then followed by decision criteria determination. A company can create a long list of suppliers where to pick suitable options according to decision criteria. From the long list is needed to filter out a short list of prospective suppliers, after the final decision will be made. Not to remember the last step of Sonmez, monitoring and evaluating the performance of selected supplier.

Nowadays, as strategic partnerships are being formed, the significance of supply chain partner is seen as a source for maintaining competitiveness, and hence can even lead to existing supplier base reduction (Vokurka *et al.* 1996). Subsequently, Vokurka *et al.* (1996) sees that the more condensed supplier registry is, the easier it is to maintain control. It has been identified that avoiding arising disruption within organization's supply chain is one of the core initiatives in strategic sourcing (Shemshadi *et al.* 2011). Seemingly, effective supplier selection can be considered as one of the key capabilities in striving for successful supply chain. As is, Cavalcante *et al.* (2019) have defined supplier selection to be a key factor in maintaining competitive advantage in supply chain, also Tsai *et al.* (2010) argued that selecting appropriate supply partner improves firm's competitiveness. It is not only about keeping up the continuous stream of supplies, but also how reliable and solid the involved suppliers are. For a manufacturing company it's highly critical that the supply streams are in and on time, as well as orders are being placed fluently. A good relationship between a buyer company and a supplier can lead to synergy benefits (Yu & Wong 2014).

Sarkis & Talluri (2002) argues that as a strategic decision supplier selection is not trivial, rather it incorporates various aspects when closing deals. They described the process of selecting supplier as an action of multi-criteria decision-making, also known as MCDM. Similarly, Amir Hossein *et al.* (2012) determined that supplier selection is also an act for sustainable business operations and thus a kind of MCDM problem. The idea is to implement a MCDM method in defining the top priorities and finding out the best suppliers based on those predefined criteria. Sometimes, choosing from numerous alternatives trade-offs are inevitable. Here, the trade-offs may relate to both the quantitative and qualitative criteria (Amir Hossein *et al.* 2012). The selection process may involve weighting certain selection criteria over the other (Gupta 2015). For example if one supplier offers materials at lower price, but is not able to arrange delivery in accordance to desired schedule, the timing could be the determinant criteria for a purchaser. The conducted research (Azadnia *et al.* 2012) proposed that to be able to solve a MCDM problem in supplier

selection one could be using methods like Analytical hierarchy Process (AHP), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), Analytic Network Process (ANP) or alike. Also Gupta (2015) presented some of these techniques and additionally names Genetic Algorithm (GA), Artificial Neural Network (ANN), Data Envelopment Analysis (DEA), and their hybrids.

As earlier mentioned, supplier discovery precedes supplier selection process. However, before any supplier can be selected, a list of prospective suppliers is required. Consequently, in comprehensive supplier selection process are potential suppliers first being identified, screened and then evaluated, following with a deeper analysis about their specifications and capabilities, and finally resulting in signing a contract. The selection process not only involves professionals from company's supply chain business unit, but as well financial management experts or even category managers may be engaged. As a multidisciplinary process selecting a supplier also demands financial and human resource efforts. Eventually as the final decision about the selection is made, the most lastly a purchase order will be placed. According to Abdul Zubar & Parthiban (2014) supplier selection sits astride of the first two stages of below presented structure:

1. Criteria for establishing the first round suppliers to be evaluated
2. Criteria for decisive supplier selection
3. Placing the purchase order and order specifications

Above structure described of how the actual purchase order gets placed and what steps go before. Clearly can be seen, that selecting a supplier is only part of the whole procurement activity. It all starts with determining the suppliers going in for a preliminary round of evaluation, so called longlist of prospective suppliers. After an evaluation and sorting phase, the most promising options will end up in the shortlist of suppliers. In the end the supplier selection "funnel" narrows down all the options, so that the outcome would represent only those who would actually receive the request for proposal.

Obviously, supplier selection requires multiple aspects to take into account. Chai *et al.* (2013) argued that when selection decision becomes more strategic and complex, even more qualitative factors are being involved (e.g. environmental sustainability). In the beginning of supplier selection history the main three criteria have been price, quality and time (Abdul Zubar & Parthiban 2014). During the past years according to Chai *et al.* (2013) also more indirect factors have been identified to have an impact on the supplier qualification like relationships and commitment, more intangible components. Ye *et al.* (2014) concluded in their study that the selection bases on financial and managerial criteria: quality, cost, delivery, and other performances. Cheraghi *et al.* (2004) provided a conclusion about the evaluation criteria for selecting supplier:

1. Quality is top most evaluating criteria followed by delivery, price and service.
2. It is found that reliability, flexibility, consistency and long-term relationship as significant new entrants of critical success factors for supplier selection.

Among the above stated aspects Cheraghi *et al.* (2004) additionally mentioned that in general the core objective is to mitigate the risks and maximize the total value for the purchasing company. Thus, also Cheraghi *et al.* (2004) required to add extensively other key performance indicators for instance as following: leadership, trust, commitment, communication, involvement, conflict resolution techniques and resources. These factors frame the models or approaches to be used when conducting the actual supplier selection.

Like Vokurka *et al.* (1996) emphasized the importance of strategic partnerships between buyer and supplier, an organization has an option to build its procurement operation foundations on either single or multiple suppliers. This balancing between whether to go with single or multiple suppliers is also considered to be part of strategic sourcing decisions (Abdul Zubar & Parthiban 2014). Abdul Zubar & Parthiban (2014) identified in their research that when an organization does not have any criteria limitations, one supplier can satisfy all the requirements defined. This is called *single sourcing*. Vice versa, when an organization faces a situation of setting various constraints for procuring the goods, multiple suppliers are needed in order to be able to fulfill the whole order capacity, ergo *multi sourcing*. This is considerably dependent on the characteristics

of the procuring company, whether it needs to decentralize purchasing to multiple sources or able to achieve all the required materials from one place only. Decentralizing could be an option, when a company seeks for better risk management, as decentralizing means also independency on one single supplier. On the other hand, decentralizing can in worst case scenario lead to too scattered supplier base and may become more difficult whole to manage. In contrary, Gupta (2015) argues that procurement manager may want to split the purchase orders between suppliers in order to creating a constant competitive characteristics between the supplying companies.

Seemingly, selecting suppliers in company supply chain operations is relatively complex and largely both strategic and operational multi-criteria decision. Multiple factors play role in choosing the best fitting supplier in regarding business processes. Supply selection has a significant impact on organization's supply chain performance and efficiency, and enables it great leverage if adjusted smoothly. Selections not always follow price over quality, nor other way around. It's about balancing between different predefined criteria and supplier capabilities, sometimes even affiliating with a sufficient and reliable partner is more advantageous and beneficial to a company than going with lowest prices. In a longer run literature implies that creating and nurturing long-lasting buyer-supplier partnerships are a recommended strategic choice, and for some occasions having a tight relationship with single partner can offer competitive advantage in fierce competition. Obviously, supplier selection sets quite a much restrictions and guidelines for supplier discovery process in strategic supply management. (Cheraghi *et al.* 2004; Shemshadi *et al.* 2011; Abdul Zubar & Parthiban, 2014; Yu & Wong, 2014)

2.3 Natural language processing

In this chapter will be addressed a fundamental framework for understanding what natural language processing is. The chapter targets to deliver a very introductory level conception about natural language processing, thus give the reader a grasp of the key terms, methods and techniques. The chapter covers briefly some prior research on the topic, its relation to such entities like artificial intelligence (AI), machine learning (ML) and deep learning (DL). Additionally, natural language processing is here linked to concept of text mining, which is relatively relevant domain in order to understand the conducted thesis research.

2.3.1 Basics

Natural language processing or henceforth *NLP* is an integrative field of study which combines computer science, artificial intelligence and cognitive psychology, ergo is concerned as an interaction happening between human language and the computer. Deng & Liu (2018) describes NLP purpose to be processing or understanding human language by using computer. This approach is also supported by Hardeniya *et al.* (2016) with a claim stating that NLP refers greatly to computation linguistics and studying language with computer applications. Beysolow (2018) in contrary entwines NLP to computer science, deep learning and machine learning, thus annotating that NLP seeks to allow computers to understand human language "naturally". Subsequently, this would mean computers understanding the sentiment of text, speech recognition, and even generating question responses.

Rapidly evolving NLP has broaden its application areas over time. According to Beysolow (2018) the roots of NLP breeds already from 1940s, when formal language theory started to develop. Nowadays capabilities of NLP have been applied to domains like speech recognition, lexical analysis, text summarization, chatbots, text tagging etc (Hardeniya *et al.* 2016; Deng & Liu 2018; Goyal *et al.* 2018). Commonly associated terms with language processing are phonetics, morphology, syntax and semantics (Goyal *et al.* 2018). When studying text set similarities especially semantics is essential term to understand. It aims to understand and examine meaning of words and how single words compose meaningful sentences.

Obviously, NLP is a consequence of human and computer. NLP utilizes the vast spectrum of applied machine learning algorithms. Like earlier mentioned, NLP examines the nuances of human language with computers, but also aims to teach them how to process. Some good

practical illustrations would be voice assistants found in today's smart phones and smart speakers, like Alexa and Siri (Bokka *et al.* 2019).

The categorizing of natural language processing may vary depending on the context and research approach. In some cases natural language processing is considered as a subcategory of machine learning, or machine learning even under the deep learning which then goes under artificial intelligence or AI (Beysolow 2018; Deng & Liu 2018). Also in some occasions a term "text mining" is heard. That will be examined briefly in the chapter 2.3.2.

2.3.2 Text mining

Now, where natural language processing referred to understand and processing human language by computer, text mining in contrary relates to the discovery and extraction of interesting, non-trivial information from unstructured text (Kao & Poteet 2007). It is, as one might think, a relatively broad subject as there exists many ways to examine and mine text, moreover plethora of different mining tools and techniques for text analysis.

Furthermore, it embraces the concepts of information retrieval, text classification and clustering. Hence, text mining can be considered as an umbrella term for bigger picture of textual data science. As natural language processing encompasses the actual deeper language analysis, the collections of texts and words, and their relations within a corpus, a body of text, text mining rather includes various techniques, stages and approaches to examine text in general. Natural language processing is seen to be subconcept, a set of activities or task execution methods, of text mining (Kao & Poteet 2007). The necessity to generally understand the "differences" between text mining and natural language processing helps grasping the research conducted.

2.3.3 Common preprocessing tasks

Generally, before proceeding into actual text processing phase, a couple of preprocessing stages are needed to go through. Here are briefly introduced the very basic stages of natural language preprocessing tasks commonly done when classifying text data.

To start with, *tokenization* or lexical analysis is the process of splitting the whole text corpus into smaller individual sequences, tokens. These tokens can in theory be for example phrases, words, single letters or other meaningful entities dependent on the implemented algorithm. During tokenization the text is broken down into single words, numbers or values. Practically, there are various amount of different tokenization techniques as there are many ways to split the text into smaller pieces. Unlike, intuitively a human reader would think tokenization may not always be executed by splitting the text down to single words only, but could be also done even on more detailed level. Bokka *et al.* (2019) for instance showed a way to break the words into n-grams, where n depicts the number of characters picked out from a word. The general idea is to analyze the given word and its different forms of occurrence.

Secondly is covered *stemming*. In linguistics stemming is a preprocessing method to return a word back to its root form. This means that stemming is performed on a corpus in order to reduce the included words to their "stem" or root form. This does not necessarily mean reducing a word to its base or dictionary form, but sometimes just to its canonical form, the natural form of a word. Exemplary, word "books" could be stemmed to "book", likewise "annoying" would be "annoy". (Bokka *et al.* 2019)

Then comes *lemmatization* which is like tokenization, lemmatization is one of the key preprocessing steps in natural language processing. The purpose of lemmatization is to perform a type of word formal reduction, so that it would obtain its root form. Commonly, lemmatization is carried out with support of WordNet, an English language database, in order to be capable to determine the root form of any known English word (Leeuwenberg *et al.* 2016). Hence, lemmatization does not only cut off the ends of a word, but also compares it to an existing library. In practice, lemmatization could perform a transformation of word "better" into its root from "good", as the first one is the comparative form of the latter, "good". Being more organized process,

lemmatization is discovered to take more time to be executed than stemming, thus it is not recommended to be carried out when handling a larger corpus. (Bokka *et al.* 2019)

2.3.4 Word vector representations

The fundamental idea behind word vectors is to represent words in mathematically rendered way in some multidimensional space, ergo depict each word with its unique vector representation, in a way that semantically similar words would have similar representations in the regarded vector space. That then allows comparing the word meanings between each other. Here is offered a grasp of this applied NLP domain complexities, when desired objective like translating and transforming raw text or a collection of documents with context into machine readable form exist.

In the world where constantly growing written communication embodies a lot of information a need for analyzing and machine automated processing of the considered text data is emerging. As the textual data may include a plethora of information, and ergo descriptions of things and concepts, the requirements for machine readable text data from the linguistic approach arises. As already mentioned earlier natural language processing represents that field of science developed not only for textual data but comprehensively all linguistic and semantic information interpretation and sophisticated machine translation. As an application of machine learning and computational techniques NLP aims to make sense out of human spoken and written text (Bhattacharjee 2018).

In computational linguistic the commonly appearing term *word-space* suggest that a word's meaning could be represented with word vectors in *n*-dimensional space. Speaking of word vectors, it is needed to understand the concept of *distributional semantics*, which refers to range of variations to represent computationally different word meanings based on the patterns of co-occurrence of words in regarding text. According to Bruni *et al.* (2014) and Sahlgren (2008) these distributional semantic models (DSM), or also known as *word embeddings*, are just a textbook example of successful practical implications of computational linguistic. Moreover, enabling to provide reliable approximations about semantic relatedness. As is, one fundamental definition for *the distributional hypothesis* by Sahlgren (2006):

> *"Words with similar distributional properties have similar meanings."*

Hence, each word in the given text can be represented by a word vectors. This means that each word has its own way to be represented in the vector space. Further, this computationally generated vector approximates the word meaning obtained by translating each word occurring in text.

In DSM, mathematically a word vector describes the patterns of co-occurrence of the word within a corpus, a real world text sample, and thus approximates the meaning of a word (Lecun *et al.* 2015). As each word has its very unique vector space representation, the similarity between other words can be quantified and measured precisely. Similarity, or more generally relatedness (Bruni *et al.* 2014), is gauged by using the vector metrics in terms of geometric distance in given vector space, indicating that the closer they are in space the similar the words are expectedly. A practical example of this word relatedness would be that weekdays (e.g. Tuesday and Wednesday) appearing in the same text corpus would they would receive quite similar word vectors. Like words "Queen" and "King", or "King" and "Man" in the same corpus, an adapted example presented in the study by (Lecun *et al.* 2015). This example has been illustrated in Figure 3. Rough estimation of the similarities between "Queen" and "King" are shown as the direction of the vectors are approximately same, as well the distance between those two vectors is quite little, likewise the vectors for "King" and "Man" possess seemingly relatively same vector length. Of course, in "reality" the vector space may be multidimensional though they are in this case represented in two-dimensional coordinates.

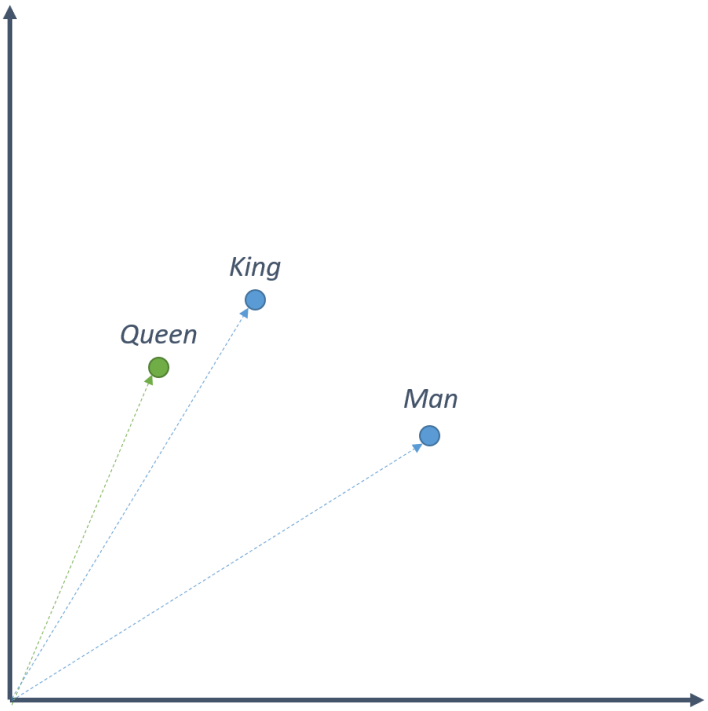

Figure 3. *Exemplary representation of word vectors for "Queen", "King" and "Man" in given vector space*

The mentioned cosine similarity simply measures the Euclidean distance between two or more vectors. The cosine similarity is determined as following:

$$\text{similarity}\ (A,B) = \ (\cos(\theta) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A depicts one word vector (e.g. for word "King"), and likewise B stands for the other (e.g. "Man). Now the angle between these two vector is represented as θ, as seen in Figure 4.

The equation is quite simple to solve with basic algebra, but as the dimensions of the vector space increases immensely it is not adequate to calculate that without using computational power. Subsequently, the cosine similarity is being calculated Figure X illustrates what the angle actually exposes.

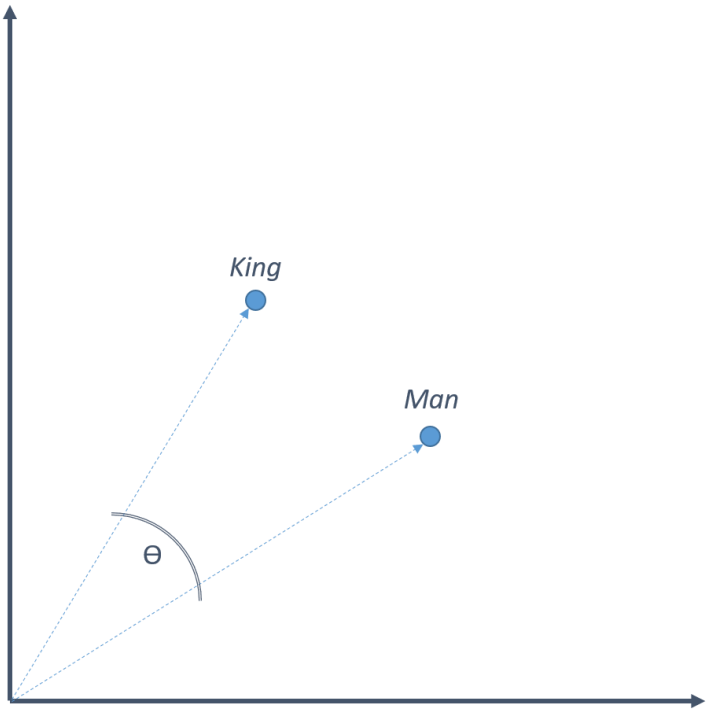

Figure 4. *Now the cosine distance between "King" and "Man" is being represented*

In the end, the distributional hypothesis is largely implemented in wide range of computational models. The distributional research indicates that word meanings could be inferred from its vector distributions across contexts. Furthermore, the DSM in general applies a co-occurrence matrix where the columns depict a concept and the horizontal rows stand for context. Then the co-occurrence frequencies are calculated for concepts and context. Note, that here the rows are n-dimensional distributional vectors, which means that if the distributional vectors turn out to be similar, the concepts occur in similar contexts. (Sahlgren 2008, Sahlgren 2005, Bruni, Tran *et al.* 2014)

The above demonstrated word embedding example is probably one the most common that exist out there. Originally illustrated by Mikolov *et al.* (2013) this distinctive model was also capable of automatically derive corresponding match between countries and their capital cities from the sample data.

2.3.5 Continuous bag-of-word (CBOW)

Before even touching fastText or Word2Vec, which are the tools from composing the word embeddings, some brief background of how do they work would be good to go through. There are actually two different ways how to compose the word embeddings or vector representations: the first one is continuous back-of-words, abbreviated as CBOW, and the latter one Skip-gram. First is introduced CBOW.

According to Mikolov *et al.* (2013) continuous bag-of-word architecture is able to predict the current word by assimilating the context. This is done by sequencing the particular word's sentence (or even the whole corpus) it appears in, and analyzing the structure of it. In the end, CBOW method predicts the current word based on the results of sequencing or "neighboring words". The order of how words appear in context does not influence the projection. The basic idea of CBOW is shown in Figure 5 with an example sentence like "the food was delightful".

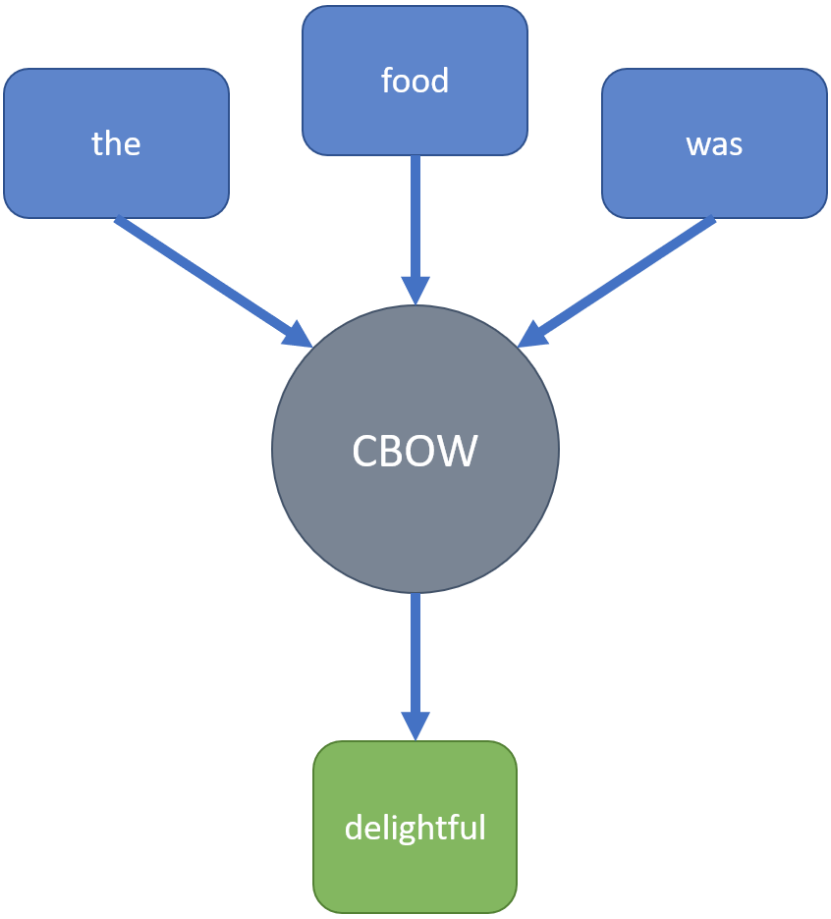

Figure 5. *Simplified illustration of CBOW algorithm operating principle*

Compared to Skip-gram method, CBOW is said to be faster and related to higher accuracy with more frequent words (Bokka *et al.* 2019). As seen from the above figure illustrating that the method is seemingly able to predict in what context a word would be probable to appear in.

2.3.6 Skip-gram

The other word embedding method is Skip-gram, which is a type of opposite of CBOW. Where CBOW was able to predict the current word depending its surroundings, Skip-gram works the other way around. The algorithm seeks to predict the surrounding words in the corpus or context based on a given input word like “delightful” in the example shown in Figure 6. Depicted below the algorithm first takes the word as input and thus generates expected neighboring words:

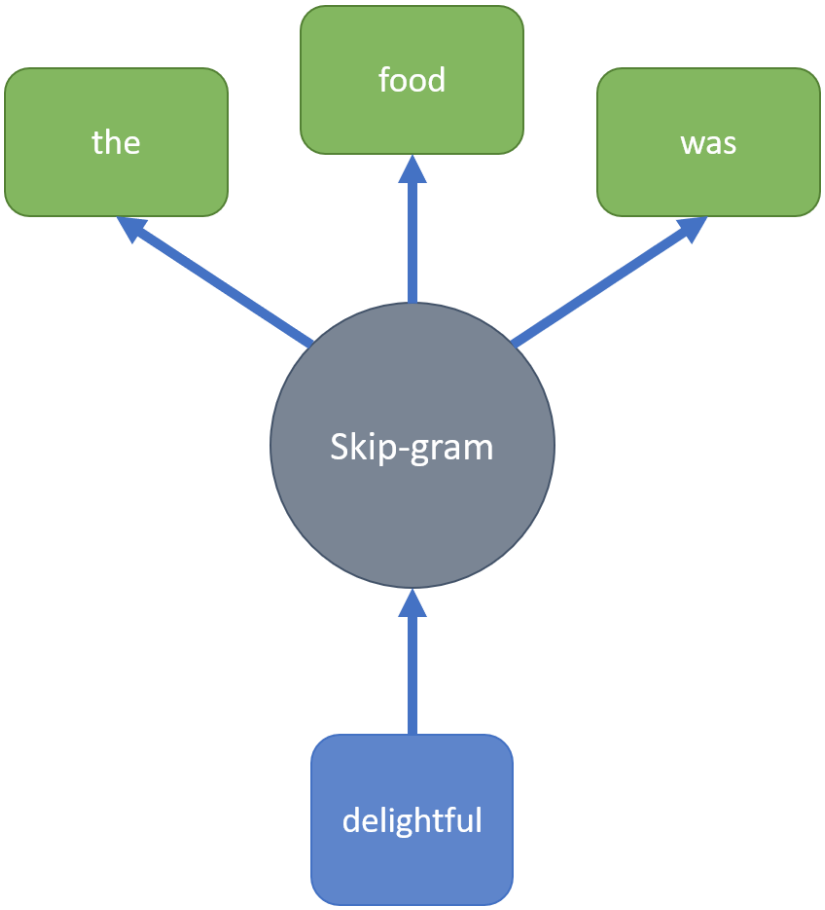

Figure 6. *Simplified illustration of Skip-gram algorithm operating principle*

This indicates that vectors are able to encode not only word similarity, but word-pairs similarities as well.
All in all, where Skip-gram methods works reversible to CBOW, as it strives to predict the surrounding words in the given context based on the current word, both the methods consider the word's surrounding in question. Despite Skip-gram might be slower to train it can outperform the CBOW in composing of word vectors both for frequent and infrequent words as well (Mikolov *et al.* 2013).

2.3.7 High-dimensionality reduction

As the dimensionality of given vector space may grow tremendously high, a need for reasonable dimension reduction arises. High dimensionality becomes problem when considered context vectors dimensionality grows. This is causality of direct function of the size of the data, thus the dimensionality increases along the word vocabulary in word-based co-occurrence (Sahlgren 2005). In this thesis dimensional reduction wasn't required, but rather makes much more convenient to illustrate spatial word presentation in human interpretable way. The key is to reduce high-dimensional data representations in a low-dimensional space. Benefits of this are e.g. reduction of the data sparseness and, obviously, dimensionality (Sahlgren 2006).

The high-dimensionality reduction can exemplary be performed by using *Principal Component Analysis* (PCA) or *t-distributed Stochastic Neighbour Embedding* t-SNE. These two are commonly used among the researches (Liu *et al.* 2018). PCA targets to recombine a linear way to express the original vector basis usually by exploiting *Singular Value Decomposition* or SVD. SVD is a matrix factorization method which composes a projection of the data in a lower dimension form from the significant singular vectors (Minhas & Singh 2017).

T-SNE in turn visualizes the high-dimensional data by pointing them a location in a lower-dimensional space. This leads to such a low-dimensional vector representations where high-dimensional vectors will be close to each other and vice-versa dissimilarity is seen as more distant vectors to each other (Van Der Maaten & Hinton 2008).

2.3.8 Generating word embeddings

In this thesis is required to understand how textual data is converted to a form so that it can be processed with computer. As earlier mentioned the word vectorizing is the very first step towards computer interpretational form. Next up, the actual word embedding generation. Computationally and mathematically composing word embeddings can be considered quite complex and multilateral concept full of elusive equations, even not necessarily demanded for a reader to assimilate.

However, the following sections will briefly shed some light on the characteristics of fastText and Word2Vec in order to understand their basic operating principles and word embedding generation theory. Mikolov *et al.* (2013) initiated the original fundamentals for word embedding algorithms and its operating principles, which is simplified in Figure 7.

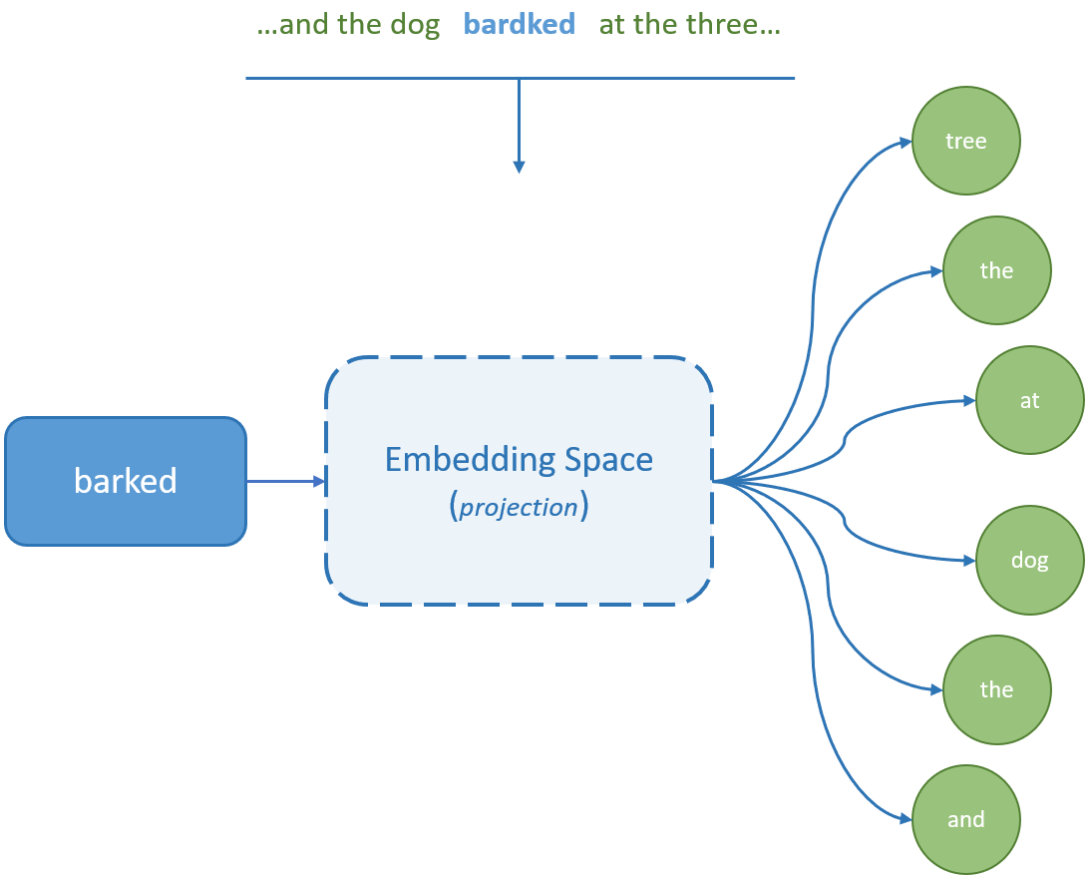

Figure 7. *Figure shows how word embedding for "barked" is related to the other words appearing in the same context*

Here, the implemented Skip-gram method predicted that the words on the far right are probably related to the word, "barked", on the left-hand side. As, in the end generating word embeddings targets to understand the context of a word, not only its appearance frequency. In the following

sections are depicted two of the most common techniques to compose word embeddings. To simply put, fastText is an extension for Word2Vec.

Word2Vec is used to compose the word embeddings by utilizing two different techniques Skip-gram and CBOW (continuous bag-of-words). Initially introduced by Mikolov *et al.* (2013) Word2Vec aims to compose word vectorization using those learning algorithms, still remaining the meaning and context of document's words. The mentioned vectorization learning algorithms are briefly covered in couple of forthcoming chapters.

According to the research by Mikolov *et al.* (2013) Word2Vec is state-of-the-art level word vectorization technique, in which a relatively simple back propagating neural network learns to vectorize when a huge datasets of words is run through. In the study the research team compared other previous models to the proposed Word2Vec model, resulting that Word2Vec outperformed. The model uses internally a simple neural network of single layer and captures the weights of the hidden layer, thus cannot be considered as a deep learning model, as only two layers exist.

The similarities between word vectors are calculated via basic cosine similarity. Word2Vector is capable to execute that simple algebraic operation and come out with vector similarity measures.

Word2Vec supports both word embedding methods, CBOW and Skip-gram. Depending on the data being processed and the use case, one might outperform over the other. Generally Skip-gram is discovered to perform well with small amounts of training data, whereas CBOW outperforms considering calculation time (Bansal & Srivastava *et al.* 2018).

FastText is quite similar to the Word2Vec, it first starts with preprocessing the body text, corpus. Then happens tokenization, where the test is divided into individual pieces or tokens. Rephrased, that while tokenization the method learns the word boundaries in the given text.

As tokenization is done fastText creates the word embeddings by using either CBOW or Skip-gram. When fastText is implemented and used in code, user can adjust the regarded hyperparameter of which one to go for. There are no right or wrong settings for the hyperparameter, rather is required to take into consideration the use case and test with different arguments (Bansal & Srivastava *et al.* 2018).

2.3.9 Clustering in NLP

In general, clustering is considered as a technique for grouping scattered objects into smaller sets, clusters. The idea is to coarsely divide those objects to belong in the same group, than the ones in the other groups. Clustering, the task of group, is not regarded as a distinctive algorithm itself, but rather an umbrella entirety embodying multiple various of algorithms capable to carry out the clustering task. In has been discovered that for exploratory data analysis tasks traditional clustering algorithms perform quite poorly. Usually, clustering algorithms need to be adjusted with hyperparameters like amount of prior clusters. This may reduce the efficiency of the clustering algorithm in question, or leave out some critical information. In order to hinder this consequence, for example elbow method is used, which is a method designed for identifying the optimum amount of clusters during the cluster analysis. In practice, the "elbow" is the state where adding another cluster won't deliver much better results out from the data being clustered. Argued as it is, the elbow method is not considered as very reliable. (Isod & Sahu 2013; Campello *et al.* 2015; Hardeniya *et al.* 2016; McInnes & Healy, 2017; Saxena *et al.* 2017)

Clustering algorithms are needed, when grouping the earlier composed word vectors and finding similarities between them. The commonly implemented clustering algorithms may be e.g. K-means and hierarchical clustering (Hardeniya *et al.* 2016). From these instances, K-means aims to find K number of groups inside the whole data population, and determine them in accordance to the mean of datapoints. First, are random datapoints chosen as the centroids of all points, and then the algorithm starts iteratively assigning the different datapoints around to its nearest centroid. During the each iteration, a recalculation of each centroid location happens, continuing so far as no centroid position changes.

The other approach for clustering, was hierarchical clustering, of that a practical options would be HDBScan or Hierarchical Density-Based Spatial Clustering of Applications with Noise. How HDBScan differs from K-means is that, whereas K-means is considered as a center-based clustering method, HDBScan instead is density-based as one might see from the whole name (Pei *et al.* 2013; Hardeniya *et al.* 2016). Density-based simply means that the clusters are being formed based on the areal density of the given data. Density-based clustering is used for irregular or intertwined, and when noise and outliers may exists in the data. HDBScan, in fact, is highly recommended technique to go with distinctively in text clustering context as it takes into account particularly the word vector densities in multidimensional space (Campello *et al.* 2015; Hardeniya *et al.* 2016; McInnes & Healy 2017; McInnes *et al.* 2017; Tahvili *et al.* 2018). Also, HDBScan considers the distances between each word vectors and comes out with a set of clusters, as well as provides a set of non-clusterable vectors. The excellence is exactly the feature, that HDBScan does not necessarily try to force every single vector to some cluster, but can leave them out, unlike other clustering algorithms (McInnes *et al.* 2018; Tahvili *et al.* 2018).

3. PRIOR INSTANCES OF NLP IN SUPPLIER DISCOVERY DOMAIN

Approaching the set research domain requires unwrapping the black box of applied NLP and text mining techniques inside the strategic supplier management context from prior research point of view. This chapter pursues to cover briefly related research in the field and conclude academic deductions based on literature review.

Assuming that plethora of supplier intelligence is sprinkled around in numerous supply chain management systems and strategic corporation ERPs. These enterprise systems comprise of massive amount of procurement data, supplier information and stores data about past transactions and parties involved. Understanding what kind of supplier intelligence this big data may encapsulate and how that data could be exploited by translating into human interpretable knowledge, would be a superior outcome from managerial point of view. Exploiting, for example natural language processing or other artificial intelligence techniques in strategic decision-making systems is not quite new practical implementation domain. Indisputably, there are various practical segments and industries where NLP based methods were used to accelerate and support decision-making process (Demner-Fushman *et al.* 2009; Carrell *et al.* 2015).

Natural language processing applications have been discovered for instance in such practical cases like analyzing textual information pursuing create predictions about e-commerce companies' probable success (Thorleuchter *et al.* 2012), a supplier selection application for manufacturing industry based on semantic analysis (Li *et al.* 2018), and also in context of global supplier selection risk assessment where text mining was exploited (Su & Chen 2018). Prior research also indicates that text analysis or semantic based techniques have been related and implemented directly to supplier discovery operations and strategies. For example in e-procurement field one research succeeded in discovering a novel way to assess applicability and compatibility of using NLP methods for mining bidding candidates from historical datasets of procurement documents and bidding applications (Aravena-Diaz *et al.* 2016).

Besides specifically NLP or semantic techniques, according to literature review also other machine learning approaches have been used in general in supplier selection domain. In a study was developed a hybrid technique based on machine learning, and concluding to increase supplier selection resilience. The research (Cavalcante *et al.* 2019) identified that manufacturers' ERP systems and databases, for example those meant for purchasing and material requirements planning purposes, preserve large volumes of data, which can then be used for risk and vulnerability assessment and predictions, and thus leverage their strategic supply chain operations. Cavalcante *et al.* (2019) introduced a supplier selection model based on supervised machine learning approach. The research conducted technically a resilient model for selecting suppliers using data-driven simulation. The notable foundations were the ability to bring contribution to a risk mitigation strategy and resilience management models, moreover for Cavalcante *et al.* there was performance data about the suppliers available to leverage the results of the research.

However, the prior research implies that NLP or applied semantic machine learning technologies are not totally novel study areas, it is still in some cases in its infancy. Multiple research have discovered the applicability of practical machine learning, particularly natural language programming, but within regarding industry though lack of data-driven organization culture. Often the challenges were identified to be data sparsity and ambiguity causing deficiencies and barriers to interpret the data in certain level of confidence. This would indicate that enriching the procurement and purchase data could be one way to leverage research.

In the light of academic literature usually how NLP is being implemented is that from the shattered data are first collected sample points which are then evaluated. In case study by (Li *et al.* 2018) was derived a semantic multi-agent systems which aimed to assist business integration. As the result they developed an architecture of how shipyard could achieve better collaboration power

within distributed manufacturing background when executing supplier selection. In the research was used a wide collection of data sourced from enterprise systems. The challenge was that the given big data, as usually, was sparse and scattered across different enterprise infrastructures.

Ameri and McArthur (2014) in turn argued the relevance of enriching semantic data in supplier discovery. The research identified the significance to enrich source data semantically in order to obtain better results in Semantic Web Rule Language method, a technique they developed. The idea behind was to apply semantic rule modelling for discover supplier intelligently. Especially manufacturing ontologies emerged from the data to play a key role in providing the required means for explicit knowledge representation.

It seems that the machine learning techniques, and NLP algorithms, performs as expected, but the barriers to gain sophisticated results could be due to available data. Vice versa, this could be put in a way that, no algorithm carries fully out from all the tasks they were initially developed because of data impurity or noise.

4. DATA

4.1 Sievo spend data

This thesis utilized real-life spend data gathered from three different Sievo clients. All of the companies were current clients representing telco industry, and hence the used data was already on premise in Sievo's database. As Sievo offers the title name software for analyzing procurement, purchase and spend data, to track down transactions, payment terms as well as for auditing supplier contract compliances, the initial idea of the concept, research topic and interest groups emerged from Sievo managerial representatives.

Before the actual spend data reaches particularly Sievo's end, and subsequently gets populated into databases, it is needed to be extracted from client's ERP. To begin with basics, the spend data contains not only such attributes as supplier name or spend amount, but also plethora of other information. Exemplary sample of containing attributes are shown below. Sometimes Sievo's client is not able to include every single attributes into spend data, hence result is much more sparse and shallow. An illustration showing the ETL-process in Sievo is provided below, in Figure 8.

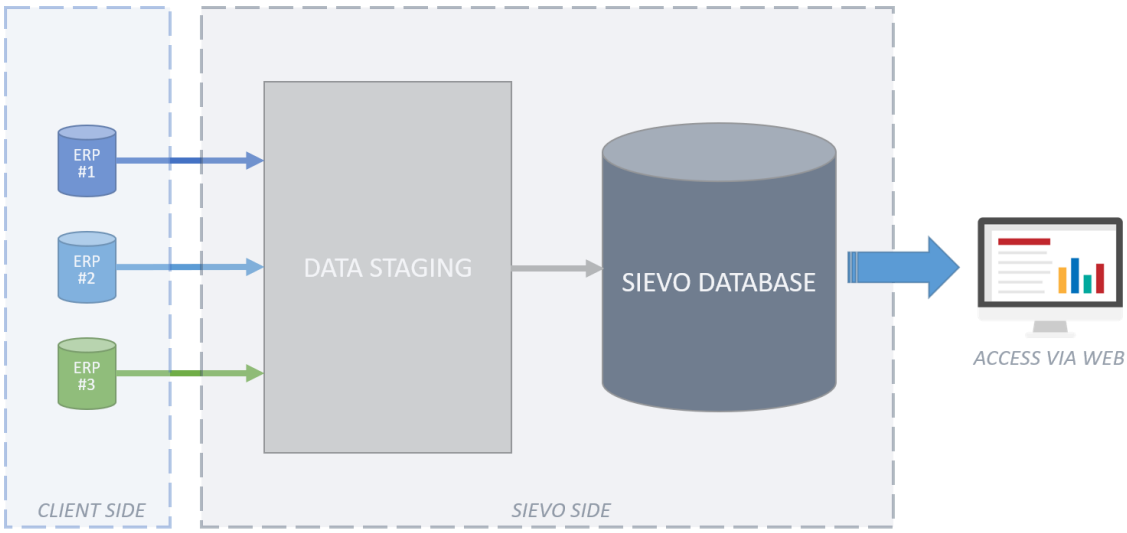

Figure 8. *Simplified ETL-process flow of spend data through Sievo system*

Figure depicts how spend data can be extracted from multiple client side systems through staging phase and finally end into Sievo database. During staging phase a data manager evaluates and checks the data correction and content. Sometimes data might have been received, but it has been delivered in wrong format or other faulty data me have been transferred. The last part, where the visual reports are published using QlikView or QlikSense, a client is able to access through a web browser. To be accurate, before spend data reaches the web browser the data does never leave actually the database side, but rather is compiled and composed as reports and then shown in graphical form through Sievo website.

Besides the earlier mentioned information pieces, there could be sometimes delivered additional attributes depending on the data batch received from the client. Client is able to configure the delivered attributes in collaboration with Sievo. Obviously, this could lead to very inconsistent data structure between different client databases. Hence, the comparison between multiple customers' spend data can be relatively challenging.

As mentioned the structure of extracted data may vary depending on the customer definition and industry nature, in addition to the spend data, some extensive data could be included like master data or market indices. This are generally delivered in different data batch, so that they will be allocated and populated in different database tables. Usually this type of data is meant for enrichment purposes of the actual spend data. Seemingly, expected that the varying data structure (an exemplary structure of data shown in Table 1) and fickle data quality would create a challenging environment for coherent data modelling, as the data sparsity and inconsistency increases.

Table 1. *An exemplary structure of spend data in Sievo database.*

DATA ATTRIBUTE	DESCRIPTION OF ATTRIBUTE
Supplier name	The supplier company name
Vendor name	Possible name of a subsidiary or alternative operation name
Supplier ID	Customer specific identification number in Sievo database
Global supplier ID	Global identification number in Sievo database
Supplier country	The origin country of the supplier
Delivery country	The country where the supplier delivered the purchase
Document line description	Commonly a coarse description of what have been purchased
Purchase order line description	Commonly a detailed version of above mentioned description. Could be even a model or serial number of considered product, item, good or material
Posting date	The date the spend data was inserted in Sievo database
Invoice date	The date the purchase was inserted in invoicing on the customer side of end
Payment terms	Describes the details of when payment is expected to due
Spend amount	The exact value of purchase (can be also negative amount, if concerns refund)
Original transaction currency	The original currency of the purchase

The spend data stored in Sievo database comprises of two different kinds of client transactions: direct and indirect. Direct spend is generated when the purchase is related “directly” to the final product, item or commodity the company produces, whereas “indirect” instead of refers to indirectly generated when purchases can be appoint to some supportive functions like human resources or buying maintenance and repair services. To simply put, indirect spend is when the purchase itself is not directly associated with producing the end product.

In this research, the initial goals was to handle only indirect type of spend as it was determined during research strategy planning to be more interesting and bring up more relevant suppliers from the noisy data. However, as the study aims to examine and conduct the applicability of natural language processing algorithms in supplier discovery, and it is supposed that the wider range of transaction data it includes, the better acting model is able to train and develop. Moreover, from the algorithmic point of view, it does not matter whether the input is indirect or direct side of spend, if they both consists of the same parameters and data points.

4.2 Sample data

One of the key concepts of this research was to evaluate underlying business opportunities in applied machine learning methods within procurement context and supplier selection process, which justified the decision that real-life procurement data was used. All the three companies have been clients of Sievo for years, which have eventually led to a massive amount of data incrementally piled up. From the research point of view, a big data pool existing consisted of a wide range of purchase information and transaction data, but no need to run through all the data while developing the model.

Due to information security reasons and sensitive nature of spend data, the regarding client company names will not be used distinctively in this thesis. Additionally, due to client data

restrictions no external data from other client sources could be incorporated in this research. Apparently these three clients in question have made an agreement of common information sharing and transparency, thus enabling Sievo to fetch and gauge their data for own product and service development purposes. In Sievo context this client cluster is known as “Telco forum”.

The used sample data was first exported from Sievo databases using Microsoft SQL Server Manager -software. It is a typical SQL infrastructure manager software for handling large datasets and querying and fetching the populated data from databases commonly used among software industry. Retrieved datasets were then stored as CSV-files on a local virtual machine inside Sievo hosted environment. CSV-format was chosen as it is popular and convenient way to handle relatively large datasets with Python. For the intended purposes of this research the sample data needed some pre-processing: constraining and cleansing due to data’s relatively sparse and sporadic nature, and moreover contained cells with empty values.

Even before actual preprocessing for purpose of natural language processing, the implemented data was imported in “consolidated” form. This means, that as the spend data structures and supplier names and identification numbers may differ between each three customers, the data needed consolidation. This was carried out already before the actual thesis research, ergo for this study purposes data already existed in suitable form. Nevertheless, it may be notable to the research.

By c*onstraining* the dataset the handling and executing became more efficient. As there was no need to run all the historical data in its entirety, the size of the dataset was also reduced. It was reasonable to optimize the size of the sample dataset to its minimum during the first rounds so that composing and executing NLP based algorithm would become more streamlined and convenient. Also expectedly the elapsing computation time will got shorter depending on the executed data amount. Eventually, as the model was trained and calibrated the dataset could be increased.

So from the fetched data only transactions posted in client side ERP between 1st of January and 31st of December 2018 were taken into account, by doing this was obtained as well consistency in data scope across the three different client. Note that the actual purchase could have been happened even earlier than on its posting date. In Sievo spend data context the posting date means the date, when it has been added to the particular system. Also while querying the sample data from Sievo database it was set to be limited to cover only the top 10 000 rows in ascending order of date. This did not only optimize the model development, but also querying time from the database instead of retrieving the entire data for year 2018 transactions.

Considering the avoidance of possible negative outliers or anomalies, the data needed to be *cleansed* from empty tuples and “NULL” values (a column that’s Boolean value is zero or data is not available). During this process empty cells and zero values were removed, so that they won’t cause unnecessary noise within the dataset or in modeling phase. In cleansing stage the dirty data was adjusted to reach more uniform, coherent and well-structured state.

In the end, more solid form was obtained which gave also better visibility over the sample data. From the Table it can be seen what information the columns provide. At this point, there was no need to reduce the dataset too much in order to avoid relevant data loss. However, if later in model calibration or practical implementation phase would seem that something irrelevant is still included that can be removed or just be ignored.

After the pre-processing the dataset seemed more comprehensible. Before proceeding into actual algorithm implementation phase the dataset structure can be modified and manipulated in many ways, ergo the original dataset would not change nor diminish. This feature enables to handle and store only the desired parts of the dataset e.g. specific columns and rows into a totally new data table. This turned out to be very helpful when starting to input the actual data through the algorithm, as not all the columns were relevant in training the model.

For the training phase some of the original dataset rows needed to be manipulated and only some of columns were selected. While observing the sample data, for example from Figure 9 it is clearly seen that some rows were identical to each other. Meaning that they are not unique type of rows.

	UltimateDesc	UltimateGSSID	POLineDesc	cluster_ids
1230	BSS	72393	Bruttolønsuddannelse Dennis Rydder Gerhard Chl...	353
8851	BSS	72393	Bruttolønsuddannelse Dennis Rydder Gerhard Chl...	353

Figure 9. *An illustration of two different PO lines, regardless the similarities of POLineDescs*

These rows might not necessarily be duplicates, but rather similar transactions happened either within the same purchase order or even scattered across the year over time. That is actually quite typical case among Sievo spend data that for example if considered direct spend there are often recurring transactions that may be very similar to each other. These transaction could include purchase order line, or PO line, descriptions like “phones” or “laptops”. However, this represents well of an example of the characteristics of given dataset. In the Chapter 5 covering the actual algorithmic implementation, the developed model is presented more detailed also the input data will be further examined. In brief, what were the interesting data columns and actually used parameters while developing the model are presented in Table 2.

Table 2. *The data structure with example real-life data after reducing unwanted spend data columns from sample data.*

UltimateDesc	UltimateGSSID	POLineDesc
Johnson Controls Denmark	*105206*	*J3507 larm call 10 2017*
Itectra	*976874*	*SFP moduler til Big Data (10G)*
Eltel	*131203*	*Små transmissionsopgaver ifbm. Q4 2017 RAN*

In contrast to preliminary research plan, where the strategy was to study the indirect spend transactions and precisely only certain spend categories, in fact not any category was excluded systematically from the dataset. Can be arguable, if concerning the NLP algorithms the more vast data spectrum the more comprehensive model is expected to obtain (Goyal *et al.* 2018).

5. ALGORITHMIC IMPLEMENTATION

First, in Chapter 4 was provided some essential information about the sample data that was exploited and used in the model development. In this chapter will be introduced comprehensively theoretical background for creating word representations and how textual data could be transformed into machine language, and finally in such form that a recommendation system is able to make sense from the given transaction data input. Also the chapter covers the practical phase of the model development and actual algorithm implementation with the sample dataset. First is presented generation of the word vector representations, ergo the word embeddings using Facebook's fastText algorithm, which will then be followed by the vector clustering by exploiting Hierarchical Density-Based Clustering method, HDBScan. After these phases will in the next chapter supervene the analysis of result extraction and performance evaluation.

5.1 Setup

The very underlying approach to this NLP aided new supplier discovery was by exploiting clients' transaction data could be analyzed that the similar transactions would refer to similar suppliers. Proceeding with that approach the dataset was based on item and product names, purchase descriptions, or other collection of words, and which then would be input through the NLP algorithm in seek of finding text based similarities between two or more suppliers. And what subsequently would lead to identifying new suppliers among the data and interpret these as possible supplier recommendations. This section handles the process flow of how the data was implemented and how algorithms were fitted with Python in Jupyter Notebook -environment.

In the beginning, the sample data was imported and manipulated in to such format and structure that the implementation of actual algorithms were relatively straightforward. The basic setup was to first create word vectors out of each and every word occurring in sample data. Only for words appearing in POLineDesc-column were taken into consideration while vectorizing the word embeddings. All the code and models were running in a browser-based version of Jupyter Notebook using Python programming language. As seen from Figure 10, the data is imported to Jupyter Notebook, where also fastText and HDBScan are being implemented and executed. After the clustering has been processed, eventually the end result is received the clustered groups.

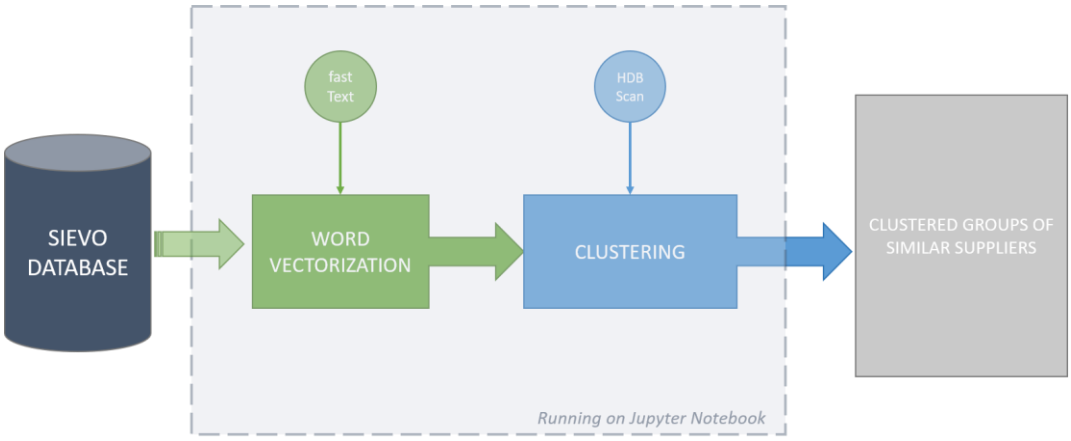

Figure 10. *Both fastText and HDBScan were running on Jupyter Notebook in browser*

In order to be able to proceed to actual word embedding composition stage, the imported .CSV data was transformed into Python data frame. A Python data frame can be then easily modified and restructured to desired form by user, and thus the data frame was transformed into a new data frame which consisted only one the columns for supplier names, their global ID-numbers and purchase order line descriptions. Consequently, the word embeddings were composed using

fastText only from the purchase order column consisting description texts, as these were the corpus where to mine the similarity measures.

5.2 Implementing fastText

As in the chapter 2.3.4 was introduced first the word representations in vector space and how they in fact mathematically are generated. In this section will be elaborated that theory part in practice and described how fastText algorithm executed the operation.

In this thesis used fastText algorithm was imported from a common Gensim Python library which core maintain organization is RARE Technologies Ltd and original author behind the development is Radim Rahûrek (Bokka *et al.* 2019). The given fastText algorithm differs a little bit from the original fastText in a way that it utilizes both CBOW and Skip-gram when creating the word vectors. By doing so it is able to compose better and more accurate spatial vectors. The Gensim fastText method was chosen due to its wide documentation and easy to implement instructions. Additionally, according to the documentation fastText should expectedly outperform compared to its concurrent Word2Vec-algorithm. Below Program 1. Shows the exact code snippet how the fastText was implemented:

```
1   from gensim.models import FastText
2
3   model_FT = FastText(size=100, windows=3, min_count=1)
4   model_FT.build_vocab(sentences=sent)
5   model_FT.train(sentences=sent,total_examples=len(sent),epochs=10)
6
7
8
```

Program 1. *Above presented a code snippet used for initializing fastText algorithm from Gensim-library.*

The given hyperparameters for the function were *size, window* and *min_count*. Size refers to feature vector dimensionality which can be predefined, window then the maximum distance between the current and predicted word in a sequence, and min_count is the set minimum value of total frequency measurement for given word to ignore it. (Bokka *et al.* 2019)

From the imported Gensim library was called also *build_vocab* function which defines the corpus initialization. It defines a variable called *sentences* as a parameter which then will be used as the sample corpus later on. Next will was called the train function which received just defined corpus “sent” as *sentences*, *total_examples* stating the count of sentences found from the corpus (in this case POLineDesc cells), and *epochs* which represents the number of iterations over the corpus, as parameters. In Figure 11 was given a sample word “Clicklab” during the model’s training phase and the composed word representation output.

```
array([ 2.3840631e-03,  2.0859309e-03,  1.1833252e-03, -2.0202361e-03,
        2.4745421e-04, -3.2752501e-05, -6.6759682e-04, -4.8318243e-04,
        5.7099038e-04, -2.1963469e-03,  3.3717910e-03,  1.3845845e-03,
        2.2072720e-03,  1.3489169e-03, -1.4157412e-03,  1.0354787e-03,
       -7.4993505e-04,  4.5303689e-04,  6.7893736e-04, -1.1621418e-03,
        7.2195340e-04,  4.0418055e-04, -2.0203149e-04,  4.1668268e-06,
        4.8281852e-04,  6.3570909e-04, -6.4978481e-04, -2.0450456e-03,
       -1.1036752e-03,  3.2570789e-04, -6.2993815e-04,  9.0575835e-04,
       -8.4423606e-04,  1.5953544e-03,  4.0621650e-03, -3.3003563e-04,
       -4.3540154e-04, -7.9444151e-05,  5.3243176e-04,  3.0390502e-04,
        1.2710338e-04, -1.2357956e-03,  6.9985696e-04,  2.1468918e-03,
       -1.6690501e-03, -7.8173174e-04, -1.2530555e-04,  8.9499209e-04,
       -1.5476590e-03,  5.0047744e-04,  7.2016753e-04, -1.9822986e-04,
        4.1733196e-04, -9.1479468e-04, -2.6104113e-04, -3.2325767e-04,
        1.0425504e-03,  8.6210360e-04, -1.2008301e-03, -6.4870052e-05,
        1.5873933e-03,  2.1486653e-03, -1.6824681e-04,  1.3301079e-03,
       -9.8983478e-04,  2.5501931e-03,  6.5325864e-04, -2.1919352e-03,
        1.0063451e-03,  1.3303569e-03, -9.4543363e-04, -1.6281295e-03,
        2.5585678e-04, -1.5305108e-03,  1.9825823e-03,  5.3430948e-04,
        6.6831149e-04, -2.0842734e-03, -5.2451657e-04, -2.2301988e-03,
       -1.5035748e-03, -1.7690018e-03,  1.1199439e-03,  3.8944141e-05,
       -5.5101345e-04,  6.8877853e-04,  4.2289929e-04,  3.4601678e-04,
        2.7288109e-04, -8.6265389e-04, -6.7009771e-04, -6.1234931e-04,
        1.1577389e-03, -8.2427979e-04, -7.9707080e-04, -9.2511834e-04,
        1.2782516e-03,  8.6517882e-04, -7.5878791e-04, -7.4606139e-04],
      dtype=float32)
```

Figure 11. *Composed word vectorization for word "Clicklab"*

As Figure 11 shows the model created vectorizations for the input word. Subsequently, the vectors for all the words in the corpus were then transmitted to a new array named "vectors", so that the title column would be "POLineDesc". Now that the newly created array consisted of all the word vectors the next step was starting to organize them into different groups based on their similarity. The similarity was based on the vectors and HDBScan run the clustering.

5.3 Clustering with HDBScan

Clustering algorithm used in this thesis was HDBScan, which stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise.. HDBScan clustering method was chosen due to remarks of (Tahvili *et al.* 2018) and as it also had a very extensive documentation and support for implementation (McInnes, Healy *et al.* 2017). According to Tahvili *et al.* (2018) HDBScan can work with high-dimensionality data much more effortless than its comparisons. Additionally regarding the sample data and its sparsity, it was expected that it could be noisy, hence HDBScan was more prominent and reasonable option due to capability to handle noise in the clustering. As already mentioned in chapter 2.3.9 HDBScan is argued to be suitable clustering technique for this thesis purposes, additionally Sievo data scientist experts supported this approach. HDBScan is often used in various different application domains such as astronomy (Low & Yang 2019), malware analysis (Morichetta & Mellia 2019) or even in accounting anomaly detection (Schreyer *et al.* 2017) or bioinformatics as well (Jia *et al.* 2017).

When the sample corpus has been transformed suitable for spatial diction and the word vectors created, next step is clustering. By clustering the vectors is able to obtain human interpretable way to "categorize" or set the vectors in the different groups based on their parameters' features. The clustering can be carried out in many different ways. There are multiple various methods and techniques which have been commonly used in scientifically research domain (Campello *et al.* 2015; Hardeniya *et al.* 2016; McInnes & Healy, 2017; McInnes *et al.* 2018; Tahvili *et al.* 2018). In the end, it does not quite matter that within what domain will clustering be used, as it's generally exploited technique in data analysis aiming to find patterns or groups in big datasets.

Frequently recurring clustering techniques includes inter alia K-means, KNN, ANN, DBScan and more sophisticated version of it, HDBScan. For high-dimensional datasets, as in this context, has

been discovered that HDBScan particularly outperforms fluently in measuring the spatial distances between word vectors and thus clusters are provided (Tahvili *et al.* 2018).

When HDBScan is used, it is first needed to import into program. In below Program 2. is first imported “hdbscan” function, which is then used for clustering. The lines 3 and 4 represents the process of appending the word vectors into an array. Later on the line 7 is defined a function called “clusterer”, which then receives the vector array as input.

```
1   import hdbscan
2
3   vectors = []
4   for i in df_reduced[‘POLineDesc’]:
5       vectors.append(model_FT[i])
6
7   clusterer = hdbscan.HDBSCAN()
8   clusterer.fit(vectors)
```

Program 2. *Depiction of HDBScan function code implemented in the research.*

In the research HDBScan clustered the sample data vectors into 374 different clusters. This happened also on the second run when the Jupyter kernel was restarted and the vectorization done all over again. It is relevant to understand that though some of the vectors may point the supplier to belong in to some distinct cluster, it may have other vectors inside another cluster. This basically means that the regarding supplier has transaction from multiple different areas, or such those POLineDescs have such general words appearing quite frequently.

5.4 Generated results

The clustered vectors were then stored in a new data frame. When testing the algorithm with one supplier specimen, a randomly selected GSSID (a distinctive identification number for a supplier) was given as input and in accordance the model started to seek matches in the data.

The fluctuation between newly discovered suppliers turned out to be significant. For example, for one test supplier the model recommended merely seven (7) relevant suppliers based on the given data, as in turn for another there were clearly over 200 matches. This large deviation between the discovered supplier recommendations would be explained due to data insufficiency or lack of descriptiveness in POLineDesc column, hence the richer the data points were in the transactions the effective the algorithm would perform, ergo even more relevant suppliers were discovered. As it was shown in Figure 9 can be clearly seen that even though the descriptions for purchase order line may be identical, the lines are different purchases (the first unnamed column separates them from each other). This means that these could belong into same purchase consignment and thus have identical purchase order line descriptions. The algorithm has obviously interpreted these two lines separately, as it can be seen from the first unnamed column where the row numbers are presented.

If taken overview of the discovered suppliers in its entirety, Figure 12 depicting the comprehensive list clearly shows how similar all the PO lines apparently were, as should.

POLineDesc	cluster_ids
Bruttolønsuddannelse Thomas Krogh (HOMA@cbb.dk...	353
Bruttolønsuddannelse Thomas Krogh (HOMA@cbb.dk...	353
Bruttolønsuddannelse Dennis Rydder Gerhard Chl...	353
Bruttolønsuddannelse Cilcila Mir Afghan - afta...	353
Bruttolønsuddannelse Andreas Langelund Jørgens...	353
Bruttolønsuddannelse Mads Skinberg - aftalenr....	353
Bruttolønsuddannelse Fredrik Tavs Yde (FYDE) -...	353
Bruttolønsuddannelse Andreas Langelund Jørgens...	353
Bruttolønsuddannelse Andreas Langelund Jørgens...	353
Bruttolønsuddannelse Mathias Romby Kvist - aft...	353
Bruttolønsuddannelse Christoffer Ulrich Larsen...	353
Bruttolønsuddannelse Andreas Langelund Jørgens...	353
Bruttolønsuddannelse Dennis Rydder Gerhard Chl...	353

Figure 12. *List of different purchases with a similar description*

As algorithm should have been performed, it has clearly interpreted the term “Bruttolonsuddannelse” to be one uniting word between these single transactions, ergo the vector word representations have been conducted correctly. And as eventually the HDBScan has been executed these suppliers resulted to end up in the same cluster number 353. Now these suppliers can be checked more closely, if the customer sees them relevant. From the actual data is also able to fetch the supplier names so that the information representation would be more convenient.

6. EVALUATION OF RESULTS

This chapter covers the analysis of conducted research results and presents a couple of reckoning feedbacks collected through customer interviews. The idea was to evaluate the supplier discovery results not only from the technical point of view, but explicitly more from the value creation driven viewpoint and practical usage approach. However, due to couple of unpredictable organizational changes and transformations within the client companies during the research project not from all the engaged clients were able to receive critical feedback, that were involved from the beginning of the research. Three companies gave their initial thoughts related to the research and these feedbacks were used to give some guidance to the project progress. Now after results have been obtained, the same clients should have been participating in delivering retrospective evaluation and rough estimations of what and how this kind of supplier discovery method would or will create value on their business processes. Only few feedback was able to request.

Starting with some technical related aspects in this research, the developed model itself performed as expectedly. The implemented fastText was able to create word vectors efficiently and relatively fast, so no extra computation power was needed. Also by tweaking the hyperparameters of fastText algorithm was kernel breakdowns avoided. For example, if dimensionality (*size* parameter) would have been raised to 300, a fatal crash was inevitable. This could have been caused due to limitations of Jupyter Notebook computational performance capacity or deficiency to handle over 100 000 lines of transaction data. Thus the optimum parameter value for dimensionality was 100. Switching other parameters' values, *window* and *min_count* did not end up changing results or accelerate calculation speed.

As clustering was done by using HDBScan the research results hugely rely on its performance to cluster high-dimensional vectors. The clustering ended up to group the vectors varyingly 370 to 380 clusters during different test runs. For each run no other factors were changed, only let the fastText calculate word vectors again. Based on those vectors HDBScan did cluster a little bit differently on each time. Significantly this did not affect the discovery results. Also the cluster sizes were relatively similar each time. The group sizes varied from 6 to 8 meaning that within each clusters the maximum of similar suppliers were 8 and minimum 1.

What was surprising, was that how big were the differences of supplier amount clustering was able to find per each tested suppliers. For example, for one tested suppliers the clustering ended up finding over 100 similar type of new suppliers, as where the other tested supplier referred only to 6 new alike. This discovery lead to an assumption that something in those suppliers' transaction data must be different. So, when randomly comparing the transaction data and PO line descriptions for the longer list of suppliers and the shorter, it could be clearly seen that the shorter the list of discovered suppliers the distinctive the descriptions were. For those where the result of discovered suppliers was approximately almost 100 different options, the PO line description contained lots of vague and indefinite collection of words that did not even referred to anything special understandable item or concept. The lines where the description was clearly written and decodable at the first glance gained the best and more definite supplier discoveries. This discovery followed the same pattern with other suppliers containing quite vague PO lines. No significant differences were found depending on whether the transaction line was in English or other foreign language. On the other hand, this was expected, as fastText seeks to find meanings and the word context within the corpus, and language should not matter.

As for couple of the suppliers the model found a huge list of similar suppliers, raised a question whether these were even relevant. It was previously mentioned that the transactions were possibly insufficient to translate the data into meaningful recommendations, but also were they sometimes not even quite related to each other. The clustering algorithm might have grouped them into same cluster only based on one or two words appearing. Thus, turned out the smaller the group the better matching new suppliers the model was able to deliver.

Only from one customer which was also engaged already in early stage interviews was able to deliver review and commentary on the obtained results. According to this interviewed client the difference between short list and a massive list of suppliers is highly significant. If a procurement officer needs to look for a new supplier or is willing to compare what other comparison suppliers there exist, too large supplier pool merely even gives any additional information. In such cases client claimed that would be more efficient just to search on web through traditional ways like Google for new suppliers. However, in the interview emerged the interest for further development of the model. They saw that if the model was able to group suppliers in a way that it offers only three to five new relevant alternatives, it would really make an impact on the decision-making process. In interview was also asked whether they would be ready to pay for such a feature as an add-on in current solution. In the light of model generated results one customer stated that, if one could measure the effectiveness of the recommendations in respect of some relevant metrics, then probably. Rephrased, if no visible or significant leverage on the core business would emerge, the value created would be relatively marginal. The customer also stated that if by enriching the source data, PO line descriptions, were able to generate better discoveries the model would be worth investing. Though, in that case the relevant supplier discoveries should have been highly guaranteed or some kind of trial period be possible in order to have an opportunity to assess the actual value.

Also, one note was that if the procurement decision-maker would be able to limit the scope of given suppliers recommended, that would increase the created information value. As is, that some of the suppliers were “too” obvious alternatives so that the given information would have been able to obtain even without the model. It was described as a too obvious recommendation, that it’s not even interesting. Based on the first round interviews where clients expressed their interest in unraveling some specific spend categories where to gain new supplier discoveries, it was no surprise that the scope of the generated supplier recommendations were a little bit “disappointment”. The customer described that even if the model was able to generate a short list of relevant supplier discoveries, a procurement professional would definitely be able to adjust the parameters the feature filters the results, meaning that for example excluding some of the most obvious results would be have been reasonable. Now from the customer point of view the model worked as a black box and delivered not really extensive value. However, they expressed also keen interest towards further development incentives, and willingness to be part of the early phase test and potential implementation group.

What was not expected, that the research raised a vast amount of information security related questions like would one be able to track down the companies given the consent to utilize the data pool, or would other be able to see detailed descriptions of other customers’ transactions and inclusive parameters (e.g. price and item).

7. DISCUSSION

In this thesis were implemented fastText and HDBScan, which both can be considered belonging to machine learning techniques. The first one is counted as an NLP algorithm under deep learning application umbrella used to create word vectors and learn sentence classification. After the vector-form word representations were composed, the actual vectors were allocated into similar groups by using HDBScan clustering method. Subsequently, the model generated a list of multiple supposed similar suppliers based on their corresponding transaction data. This chapter aims to provide in-depth conclusion about the achieved research results and reflect them to the prior research in the field. Furthermore, the chapter seeks to understand the limitations and probable deficiencies in the conducted research, as well as examining the accomplishments from theoretical and practical point of view. Here will be discussed how the study succeeded in set questions and planned strategy, as well as what circumstances and factors could have been taken into consideration while conducting the study and in project execution.

The first research question was: *Can purchase order descriptions be used as source for natural language processing based supplier discovery?* According to the Chapter 6 where the results were analyzed, it can be clearly stated that textual data enables creating meaningful supplier discoveries through natural language processing technique. The text purchase orders contain offers sufficient data source for identifying new suppliers from the spend data applying NLP. As the customer also stated, they can verify the generated results as successful research goal. Moreover the research evidence supports the assumption that implemented NLP technique, fastText (and clustering method HDBScan) were able to create meaningful information. However, the quality level of meaningfulness and value is arguable. As customer stated, the supplier discoveries themselves are clearly visible, but the information the results offer are relatively inadequate in order to make robust and secure decisions either in supplier discovery or selection processes. As a conclusion, it can be stated that purchase order descriptions can be used as source for NLP based supplier discovery. The relevance of the identified suppliers, in turn, may be arguable. That is discussed later in this chapter.

The above addressed discussion partly grasped the second research question as well, which was following: *How natural language processing based supplier discovery performs in bringing value in supplier selection process?* Reflecting on research results, answering the question may be quite controversial. A fact is that the model is able to find similar suppliers among the sample data by comparing the PO line descriptions. On the other hand, that can also be done by comparing manually one by one every single purchase row, but would not be efficient and time-consuming. However, the model succeeds in identifying the comparable suppliers fast and efficiently. The actual value can be seen to be created when new related suppliers can be easily found through available text data. Also, if one customer only has access to its own data and supplier base (as it currently is within Sievo customer base), the model here offers an extensive access to the data of one's competitors or companies from the same industry. If the purchase information would be available across-industries and was multidisciplinary by nature, that would be extensive information for procurement professionals, and hence bring value to the process. Of course the data used in this research was transparently shared and all the companies have given their consent to exploit the data openly, but thinking about further implementation objectives this kind of natural language processing based modelling would definitely bring up new information for the decision-makers to use. Fundamentally, this NLP based supplier discovery performs satisfactorily in bringing new knowledge, however further development is needed, if the performance level of information generation is wished to leverage.

Also if considered the customer given feedback clearly some new information is brought up and available after model implementation. However, even though by using NLP would be able create new information the degree of the value it creates to the supplier discovery process can be a bit arguable. One factor to be taken into consideration while discovering new suppliers was price. Obviously, the price tag is included in transaction data and more closely with the purchase order, but the thing is that this information was not part of the model. In other words, the price attribute was excluded from the text processing part and hence cannot affect the generated results.

Besides the price, also no further information about the supplier references cannot be derived from the purchase order descriptions. It was noted in the literature part that procurement officers pay attention also to previous references. If for example the supplier had a list of other successful customer stories or prominent track record of deliveries, that would offer the decision-maker a lot more knowledge about the supplier reliability. In customer interview also emerged the lack of sufficient information regarding the benchmarking of identified suppliers. According to the customer, finding new suppliers is trivially interesting and could in some cases give extensive perspective, but in the end with that information cannot really proceed on in the supplier selection process. A procurement officer can always approach the supplier afterwards as the one has been identified to potentially fulfill the customer company's need, but what if that information was already available during the discovery phase. On the other hand, this stands against the fact that no sensitive data (e.g. purchase prices) was expressed not to be willing to share during the interview. Ameri & McArthur (2014) discovered that the less information about the suppliers' capabilities is available the more irrelevant the discovery results are. This approach is also supported by Lee *et al.* (2011) and Lee *et al.* (2013), as they state that supplier capabilities are often the preferred "first round" information desired to be collected

However, in the concerned study by Lee *et al.* (2011) was used string-type information, ergo plain text, to derive similarity measures between suppliers' capabilities descriptions. These were then used to track down new suppliers, supplier discoveries. Although, that research did differ from this study in respect to source data format, vectorization method (VSM) and clustering technique (PAM and graph theory-based clustering). Why utilizing the recently described method would be disputed, is that the source format was in XML form and consisted of significantly more descriptive data, as it was from internal supplier registry. Obviously, this might indicate that the source data should be richer and include clear descriptions of what has been purchased, rather than sequences of unordered numbers and characters. Additionally, what should not be ignored considering this, that this thesis used transaction data whereas the Lee *et al.* (2011) handled with suppliers' capability documents.

It seems quite trivial, that enriching the transaction data (explicitly purchase order descriptions) would the model have generated supposedly better results. That it was identified that the in cases of successful and relevant supplier discoveries the PO line descriptions were adequate, that would imply enriching the transaction data. However, as a customer referred that it would require resources to enrich the data, or even sometimes fully replace the line description and revise the purchase order fulfilling process. Arguable would be that whether a company is able to gain sophisticated level of descriptive transaction data and still retain the resources and labor costs required to make purchase orders. It would have been interesting to test, if the amount of input data changed the level of supplier discoveries. Presumably the richer vocabulary would have created better results according to Ameri & McArthur (2014) and Lee *et al.* (2011). The above mentioned implies that the resolution for *what factors affects the feasibility of using purchase order descriptions for text based supplier discovery*, can be crystallized as the degree of data richness and descriptiveness of the PO lines. Moreover, the meaningful words and the context they appear seem to have an effect on the feasibility.

In the beginning of the research couple of the main incentives were to find out how value can be created in supplier selection process by accelerating it with NLP based supplier discovery: *how business value is created* and *what factors affects the value creation in supplier discovery and supplier selection process?* Now, that is answered through the customer interview and prior research. Basically, value is been created whenever something valuable information is brought up or generated. In this study was determined that value was created exemplary when new suppliers were identified efficiently. That offers a more comprehensive view to the field of suppliers delivering relevant goods. The bigger amount and detailed information there's available, the more value it'll bring to the decision-making process.

Seemingly, more detailed source data and clustering could have had leveraging impact on the supplier discovery results. Furthermore, understanding that one supplier may deliver products and items considered both as direct and indirect type of categories at the same time, which would thus mix up the different spend types with each other. Now, if the source data would have been imported only from, let's say, indirect side of spend for even detailed commodity category, that might have enhanced the results. That though is more related to information relevancy. Though,

information relevancy is highly related to the value creation, as the more relevant the information the higher value it delivers. In contrary limiting the sample data characteristics the research would have narrowed the scope of the study quite a bit, and it could have affected the generation of sufficient results. Regarding further research, a comparative execution runs would be interesting to carry out.

Furthermore, considering a wider range of tested algorithms might have been worthwhile, or at least to make a "sanity-check" if fastText is even an efficient method. Also Word2Vector method could have been tested, which was initially replaced with fastText based on literature. One of the sub-questions in this research was: *what alternatives are there to the algorithms used in this research?* Taking into consideration as a prominent alternative to fastText, Word2Vec is at least developed and implemented in Google's services. Additionally, also Zalando has created its own choice, Flair. Word2Vec and Flair both offers similar kind of NLP library like fastText, consisting of flexible algorithms ready to be implemented e.g. in text embedding purposes. Obviously, no further guarantee of these two techniques' better performance exists, but would be interesting to benchmark the results between all three alternatives. It has been identified that Flair performs way better than Word2Vec when speaking of context-based, not word-based representations (Joshi *et al.* 2019).

Regarding the clustering algorithm alternatives, for HDBScan there are also multiple substitutes: K-Means, KNN and ANN (Pei *et al.* 2013). Understanding the algorithms behind the results, fastText and HDBScan, one could have run the data through different kind of clustering algorithm and vectorizing techniques. Even though choosing fastText was argued as the most suitable method, as well as HDBScan was too, there's no doubt that other techniques wouldn't generate different results. However, that does not mean they would have outperformed the applied ones. Also what comes to test data, the implemented dataset, the results largely relied on the data descriptiveness and how rich the PO line descriptions were. When examining the purchases, it is quite usual and common that the lines do not always contain any sufficient information at all of what the purchase is about, not to mention that text mining algorithm would be able to mine meaningful information out of it.

There are several paths to go for future research in the context of supplier discovery and applied natural language processing methods, even more avenues to be discovered regarding these two fields of science combined. A lot could be discussed, simulated and further studied, and thus enable to yield detailed information and deeper knowledge about the topics in question. Understanding the limitations of fastText, HDBScan and complexities of supplier discovery process would imply that there's plenty of rocks to be turned over for this motive. From the customer perspective the future research could take a place in better understanding what if and how much would PO line descriptions enriching yield to the results of supplier discovery.

Seemingly, a huge potential could be brewed out from the foundations of combining applied NLP and strategic supplier management context, moreover this does not comprise of only the supplier discovery domain, but also other business processes related. How about utilizing NLP in payment terms tracking and supplier compliance benchmarking, or what if text mining could contribute creating value in supplier risk evaluation and capabilities determination, not to mention deep-dive into applied methods, fastText and HDBScan. How many combinations of word embedding techniques and clustering algorithms there might be within the foundation of artificial intelligence, or machine learning, or deep learning?

Though, the application was able to perform promisingly and was capable to deliver some listing of potential suppliers found from the transaction data, that does not mean necessarily that procurement specialist would be able to make the call alone in respect to that intel. As Sonmez (2006) argued in the beginning of the chapter 2.2 the supplier selection process is quite complex and requires multiple steps. Using an application like described in this thesis would not quite be able to deliver sufficiently information about all the criteria mentioned in the background theory. Prominently for example considered risk that's related to a supplier cannot be assessed in anyway by interpreting the information retrieved from the NLP derivatives. Or at least, with the current implementation. If in future, this kind of application will be taken into use, one might think how or if in any way would be able to include risk assessment aspect into algorithm. Like earlier mentioned, for comprehensive and strategically adequate supplier evaluation is needed more

specific information about the supplier performance in previous cases, probably even other customer references is needed.

That also relates to another aspect considering the MCDM technique. As supplier selection was clearly identified to be such a decision-making problem, that multiple factors are needed to take into account, such factors that cannot be evaluated through the information received from the NLP application. Also at Sievo it is commonly discovered that by creating supplier affiliate contracts, companies are able to receive better deals and spend savings. Meaning that possibility to obtain synergy benefits exists. Incentive for changing or switch from the current supplier base, highly relies on the how vast information and deep knowledge the NLP algorithm would be able to deliver. Seemingly, at least so far, not quite capable. Simply to put, there's a lot more to do further research in that regard, so that the application would be able to more rich and such knowledge that a procurement expert can honestly make decisions based alone the application. Now, it can only perform as a support machine for finding "hidden" potential suppliers using comparison companies transaction data and analyze them in accordance to the expected similar purchased items or goods, thus this NLP application is like a "prospective generator".

One of the major shortcoming is the absence of proper model performance evaluation and comparison to other alternatives. Now, this research scope was limited in the way that no further analysis of different options for fastText or HDBScan weren't implemented, though it would inevitable to try different prototypes of different NLP algorithms in order to test and find out the most prominent and best performing one. Considering the future research topics, one could run the same sample data through different text classifier and word embedding algorithms, and then compare the generated results to fastText. As well the exactly same could have been done regarding HDBScan. Even though both the methods were highly recommended by experts, a robust extensive academical research would have definitely have some comparison and benchmarking covered. For clustering, instead of using HDBScan, one may also try implementing mentioned K-means, KNN or ANN.

REFERENCES

Abdul Zubar & Parthiban. (2014). Analysis of supplier selection methods through conceptual module and empirical study. International Journal of Logistics Systems and Management, 18(1), 72-99.

Ameri, F. & McArthur, C. (2014). Semantic rule modelling for intelligent supplier discovery. International Journal of Computer Integrated Manufacturing, 27(6), 570-590.

Aravena-Diaz, V., Gacitua, R., Astudillo, H. & Gayo, J.E. (2016). Identifying potential suppliers for competitive bidding using Latent Semantic Analysis. 2016 XLII Latin American Computing Conference (CLEI), 1-12.

Azadnia, A.H., Saman, M.Z.M., Wong, K.Y., Ghadimi, P. & Zakuan, N. (2012). Sustainable supplier selection based on self-organizing Map Neural Network and Multi Criteria Decision Making Approaches. Procedia Social and Behavioral Sciences, 65, 879-884.

Bansal, B & Srivastava, S. (2018). Sentiment classification of online consumer reviews using word vector representations. Procedia Computer Science, 132, 1147-1153.

Beysolow II, T. (2018). Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing. 1 edn. Berkeley, CA: Apress L.P.

Bhattacharjee, J. (2018). fastText Quick Start Guide. Packt Publishing.

Bokka, K.R., Hora, S., Jain, T. & Wambugu, M. (2019). Deep Learning for Natural Language Processing. Packt Publishing.

Bruni, E., Tran, N.K. & Baroni, M. (2014). Multimodal distributional semantics. Journal of Artificial Intelligence Research, 49, 1-47.

Campello, R.J.G.B., Moulavi, D., Zimek, A. & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(1), 1-51.

Carrell, D.S., Cronkite, D., Palmer, R.E., Saunders, K., Gross, D.E., Masters, E.T., Hylan, T.R. & Von Korff, M. (2015). Using natural language processing to identify problem usage of prescription opioids. International Journal of Medical Informatics, 84(12).

Cavalcante, I.M., Frazzon, E.M., Forcellini, F.A. & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. International Journal of Information Management, 49, 86-97.

Chai, J. & Ngai, E.W.T. (2019). Decision-making techniques in supplier selection: Recent accomplishments and what lies ahead. Expert Systems with Applications, 140, 112903.

Chai, J., Liu, J.N.K. & Ngai, E.W.T. (2013). Application of decision-making techniques in supplier selection: A systematic review of literature. Expert Systems with Applications, 40(10), 3872-3885.

Cheraghi, S.H., Dadashzadeh, M. & Subramanian, M. (2004). Critical Success Factors For Supplier Selection: An Update, 20(2).

Chopra, D., Marthur, I. & Joshi, N. (2016). Mastering Natural Language Processing with Python. Birmingham, UK, Packt Publishing.

Demner-Fushman, D., Chapman, W.W. & McDonald, C.J. (2009). What can natural language processing do for clinical decision support?, Journal of Biomedical Informatics, 42(5), 760-772.

Deng, L & Liu, Y. (2018). Deep Learning in Natural Language Processing. Singapore, Springer. Available: < https://www.springer.com/gp/book/9789811052088> (Cited 15.11.2019)

Goyal, P., Pandey, S. & Jain, K. (2018). Deep Learning for Natural Language Processing: Creating Neural Networks with Python, 1st, 1 edn. Berkeley, CA: Apress.

Gupta, M. (2015). Supplier selection using artificial neural network and genetic algorithm. International Journal of Indian Culture and Business Management, 11, 457-472.

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N. & Mathur, J. (2016). Natural Language Processing: Python and NLTK. Birmingham, UK: Packt Publishing.

Isod, S.R. & Sahu, A.M. (2013). Clustering Techniques. International Journal of Advanced Research in Computer Science, 4(6).

Jia, G., Preussner, J., Guenther, S., Yuan, X., Yekelchyk, M., Kuenne, C., Looso, M., Zhou, Y. & Braun, T. (2017). Single-cell transcriptional regulations and accessible chromatin landscape of cell fate decisions in early heart development, Nature Communications.

Joshi, A., Karimi, S., Sparks, R., Paris, C. & Macintyre, C.R. (2019). A Comparison of Word-based and Context-based Representations for Classification Problems in Health Informatics. Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, 135-141.

Kang, Y., Kim, J. & Peng, Y. (2011). Extensible Dynamic Form Approach for Supplier Discovery. Proceedings of the IEEE International Conference on Information Reuse and Integration, IRI, 3-5 August, Las Vegas, Nevada, USA.

Kao, A. & Poteet, S.R. (2007). Natural Language Processing and Text Mining. London: Springer.

Lecun, Y., Bengio, Y. & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

Lee et al. (2011). A Supplier Discovery Framework for Effective and Efficient Configuration of a Supply Chain. International Journal of Industrial Engineering: Theory Applications and Practice, 18(3), 109-119.

Leeuwenberg, A., Vela, M., Dehdari, J. & Van Genabith, J. (2016). A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. The Prague Bulleting of Mathematical Linguistics, 105(1), 111-142.

Li, J., Sun, M., Han, D. Wu, X., Yang, B., Mao, X. & Zhou, Q. (2018). Semantic multi-agent system to assist business integration: an application on supplier selection for shipbuilding yards. Computers in Industry, 96, 10-25.

Liu, S., Bremer, P-T., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y. & Pasucci, V. (2018). Visual Exploration of Semantic relationships in Neural Word Embeddings. IEE Transaction on Visualization and Computer Graphics, 99.

Low, J. & Yang, A. (2019). Clustering of hotspots in the cosmic microwave background. The European Physical Journal Conferences, 206(11).

Luan, J., Yao, Z., Zhao, F. & Song, X. (2019). A novel method to solve supplier selection problem: Hybrid algorithm of genetic algorithm and ant colony optimization. Mathematics and Computers in Simulation, 156.

McInnes, L. & Healy, J. (2017). Accelerated Hierarchical Density Clustering. IEEE International Conference on Data Mining Workshops (ICDMW), 33-32.

McInnes, L., Healy, J. & Astels, S. (2017). HDBScan: Hierarchical Density Based Clustering, The Journal of Open Source, Sotware, 2(11).

Mesmer, L. & Olewnik, A. (2018). Enabling Supplier Discovery Through a Part-focused manufacturing process ontology. International Journal of Computer Integrated Manufacturing, 31(1), 87-100.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Available: <https://arxiv.org/abs/1301.3781> (Cited 27.10.2019).

Minhas, J. & Singh, B. (2017). A comparative study of SVD, PCA and Clustering Techniques of Copy Move Forgery. International Journal of Advanced Research in Computer Science, 8(4).

Morichetta, A. & Mellia, M. (2019). Clustering and evolutionary approach for longitudinal web traffic analysis. Performance Evaluation, 135.

Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Doctorla dissertation, Kista: Department of Linguistics Stockholm University, Stockholm, Sweden. Available: <http://eprints.sics.se/437/1/TheWordSpaceModel.pdf> (Cited 25.10.2019).

Sahlgren, M. (2008). The distributional hypothesis. Rivista di Linguistica 20(1), 33-53.

Sarkis, J. & Talluri, S. (2002). A model for quantifying strategic supplier selection: Journal of Supply Chain Management, 38(1), 18-28.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A. & Reimer, B. (2017). Detection of Anomalies in Large Scale Accounting Data Using Deep Autoencoder Networks. Available: < https://arxiv.org/abs/1709.05254> (Cited 27.10.2019).

Shemshadi, A., Toreihi, M., Shirazi, H. & Tarokh, M. (2011). Supplier selection based on supplier risk: an ANP and fuzzy TOPSIS approach. The Journal of Mathematics and Computer Science, 22, 111-121.

About Sievo. (2019). Available: < https://sievo.com/company/about-sievo> (Cited 21.11.2019).

Sonmez, M. (2006). Review and critique of supplier selection process and practices. Available: < https://repository.lboro.ac.uk/articles/Review_and_critique_of_supplier_selection_process_and_practices/9494939> (Cited 21.11.2019).

Su, C. & Chen, Y. (2018). Risk assessment for global supplier selection using text mining. Computers and Electrical Engineering, 68, 140-155.

Tahvili, S., Hatvani, L., Felderer, M., Afzal, W., Saadatmand, M. & Bohlin, M. (2018). Cluster-Based Text Scheduling Strategies Using Semantic Relationships between Test Specifications. Proceedings of the 5th International Workshop on requirements engineering and testing, June, Gothenburg, Sweden. Available: <https://ieeexplore-ieee-org.libproxy.tuni.fi/document/8444116>. (Cited 20.11.2019).

Thorleuchter, D. & Van den Poel, D. (2012). Predicting e-commerce company success by mining the text of its publicly-accessible website. Expert Systems With Applications, 39(17), 13026-13034.

Tsai, Y.L., Yang, Y.J. & Lin, C. (2010). A dynamic decision approach for supplier selection using ant colony system. Expert Systems With Applications, 37(12), 8313-8321.

Van Der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579-2625.

Vokurka, R.J., Choobineh, J. & Vadi, L. (1996). A prototype expert system for the evaluation and selection of potential suppliers. International Journal of Operations & Production Management, 16(12), 106-127.

Wetzstein, A., Hartmann, E., Benton jr., W.C. & Hohenstein, N. (2016). A systematic assessment of supplier selection literature - State-of-the-art and future scope. International Journal of Production Economics, 182, 304-323.

Ye, Y., Jankovic, M., Kremer, G.E. & Bocquet, J. (2014). Managing uncertainty in potential supplier identification. Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM, 28(4), 339-351.

Yu, C. & Wong, T.N. (2014). A supplier pre-selection model for multiple products with synergy effect. International Journal of Production Research, 52(17), 5206-5222.