Tampere University

Noora Raitanen

NATURAL LANGUAGE PROCESSING FOR SEMI-AUTOMATED INSIGHT DISCOVERY FROM PUBLIC DOCUMENTS OF COMPANIES

A topic modeling approach

Master of Science Thesis
Faculty of Information Technology and Communications
Examiners: Jyrki Nummenmaa
Okko Räsänen
October 2022

ABSTRACT

Noora Raitanen: Natural language processing for semi-automated insight discovery from public documents of companies
Master of Science Thesis
Tampere University
Information Technology
October 2022

A consistent and successful sales process is a key requirement in consulting environment, where the work is project-based and dependant on incoming customer engagements. Sales in a highly-specified technical field require knowledge on both the sold technology and the potential customer. The amount of information available on both is vast and can be difficult to get a grasp on in a short time. Hence, the usage of different analysis tools is a common practice during the sales process. The focus of such tools is often in lead acquisition and qualification rather than in customer approach and conversation starting, which is the focus of this thesis work. Here the focus is on discovering if an automated insight discovery can provide concrete talking points to be discussed and presented to a potential customer during the sales process. The aim is to support the customer approach with additional insights and potential areas they may need support with based on the customers' public appearance

A test framework to analyze document published by companies and discover the key topics discussed in the material was built to study the research question. The framework includes separate preprocessing and analysis steps. The preprocessing runs a preprocessing sequence to the material and formats it for the analysis. The following analysis step includes the n-gram search and topic modeling. The chosen topic modeling method is Latent Dirichlet Allocation. The used source material consists of documents published by selected companies: job postings, Capital Market Day materials, and press releases. The source material is sectioned into two separate data sets. The *Company* - dataset including the documents of an individual company and the *Sector* - dataset including the documents of multiple companies from a selected business segment. The first is used to obtain insight on the talking points and focuses of a single company and the latter creates an overall view of the general focus of the whole business area.

The work conducted was assessed in two different ways: the usability of the system and the relevance of the results during the sales process. The assessment was conducted as free-form discussions with experts. The system built was overall easy to use and suitable for users without extensive technical knowledge. The n-gram search provided simple, yet effective, metric to assess the most frequent concepts in the material. The simplicity of the approach yielded results easy to understand and utilize. The topic modeling results were more complex to interpret and required additional human analysis after the initial process. The results often lacked coherence and were difficult to interpret under a theme during human assessment. Hence, the topic modeling could provide actionable insight in a limited number of cases. Overall, the analysis pipline and its results were found to be useful for customer approach in the cases where the results were coherent enough for human to interpret.

Keywords: natural language processing, topic modeling, B2B-sales, Latent Dirichlet Allocation, text mining

TIIVISTELMÄ

Noora Raitanen: Puoliautomaattinen tiedonlouhinta luonnollisen kielen käsittelyn avulla yritysten julkisista dokumenteista
Diplomityö
Tampereen yliopisto
Tietotekniikka
Lokakuu 2022

Jatkuva, onnistunut myynti on tärkeä kriteeri konsultoinnissa, jossa yrityksen tulos riippuu myydyistä asiakasprojekteista. Myyntityö teknisessä, pitkälle erikoistuneessa ympäristössä vaatii myyjältä ymmärrystä sekä myytävästä teknologiasta että asiakkaasta. Informaatiota molemmista on saatavilla paljon ja helposti, mutta suurta tiedon määrää on vaikea käsitellä tehokkaasti hektisessä myyntityössä, jolloin halutun informaation löytäminen vaikeutuu. Erilaisten analyyttisten työkalujen käyttö tiedonhallinnassa ja -hankinnassa on tyypillistä myyntiprosessille. Yleensä analyysityökalujen käyttö keskittyy potentiaalisten asiakkaiden etsimiseen ja kartoitukseen. Tässä työssä keskitytään perusteluvaiheen työn tukemiseen. Perusteluvaiheessa asiakas kontaktoidaan ensimmäistä kertaa ja asiakkaan kanssa aloitetaan keskustelu mahdollisista myytävistä tuotteista tai palveluista. Työn tavoitteena on luoda puoliautomaattinen työkalu, joka tuottaa tietoa potentiaalisesta asiakasyrityksestä keskustelun aloituksen avuksi yrityksen julkaisemien tekstidokumenttien pohjalta ja tutkia työkalun vaikutuksia myyntityön onnistumiseen.

Alkuperäisen kysymyksen testaaminen suoritettiin rakentamalla puoliautomaattinen järjestelmä, joka esiprosessoi lähdemateriaalin ja suorittaa aineiston analyysin. Analyysi sisälsi n-gram haun ja aihemallinnuksen. Aihemallinnuksessa käytettiin LDA - menetelmää (*engl. Latent Dirichlet Allocation*). Työn lähdemateriaali koostuu valittujen yritysten julkaisemista dokumenteista: Pääomamarkkinapäivien (*engl. Capital Market Days, CMD*) esityksistä ja litteroiduista puheista, työpaikkailmoituksista ja lehdistötiedotteista. Lähdemateriaalien pohjalta luotiin kaksi datasettiä. Company - aineisto (*suom. yritys*) sisältää yhden valitun asiakasyrityksen julkaisemia dokumentteja. Aineistoa käytettiin tietyn yrityksen keskeisimpien puheenaiheiden ja materiaalin pohjalta tärkeimpien aiheiden löytämiseen. Sector - aineisto (*suom. markkinasegmentti*) sisältää valitun markkinasegmentin usean yrityksen dokumentteja ja sen avulla luodaan kokonaiskuva tietyn markkinan toiminnasta.

Työ arvioitiin kahdella eri kriteerillä: Järjestelmän käytettävyys ja saatujen tulosten sopivuus myyntityön tukemiseen. Arviointi suoritettiin vapaamuotoisissa keskusteluissa aihealueen asiantuntujoiden kanssa. Rakennettu järjestelmä koettiin melko selkeäkäyttöiseksi ilman laajaa teknistä osaamista ja sen katsottiin tarjoavan suoraviivainen tapa analyysien tuottamiseen. Analyysimenetelmien kohdalla n-gram haun tulokset koettiin selkeiksi ja ymmärrettäviksi tavaksi esittää tekstissä eniten toistuvia termejä. Aihemallinnuksen tulokset vaativat jatkoanalysointia ihmisen toimiesta ja osa mallin tuottamista aiheluokista oli vaikea luokitella ihmisen toimesta mihinkään tiettyyn aihekategoriaan. Tämäm vaikeutti aihemallinnuksen käyttöä ja vähensi sen tuomaa tukea myyntiprosessiin. Kokonaisuudessa analyysityökalu nähtiin hyvänä tapana tuottaa myyntityötä tukevaa tietoa, mutta tulosten yleinen epäselvyys ja vaikeatulkintaisuus haittasivat niiden käytettävyyttä erityisesti aihemallinnuksen kohdalla.

Avainsanat: luonnollisen kielen käsittely, aihemallinnus, B2B-myynti, Latent Dirichlet Allocation, tekstinlouhinta

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

PREFACE

I'd like to thank Futurice, which of this thesis was done in collaboration. I would like to sincerely thank them for the thesis opportunity and guidance provided. Thanks are addressed to Tuomas, who provided the topic of my Master's thesis and support during the process, and to Teemu, who helped me to tackle technical challenges and supported me with during the technical development process. The thesis allowed me to have an opportunity to gain a better understanding of the sales process and needs, as well as obtain deeper knowledge on text based topic analysis.

I wish to also thank my supervisors from Tampere University: Associate Professor Okko Räsänen and Professor Jyrki Nummenmaa. Thank you for the guidance to shape to thesis. Thank you for providing comments and supervision during the whole thesis process.

Finally, and mostly, I want to sincerely thank all my friends, who have made my time at Tampere University fun and memorable, but also made the studies easier. Special thanks to my home guild TiTe and Uranaisten Opiskelijaseurafor making the years memorable. I have greatly enjoyed the past years at the university and learned a lot, but also gotten a lot of great experiences and memories outside the studies. Lastly, I wish to thank my family for the never-ending support and faith in me during my studies and the long-lasting thesis process. It took a while, but finally here we are.

Tampere, 25th October 2022

Noora Raitanen

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

α	Document-topic density
β	Topic-word density
B2B	Business to Business
CMD	Capital Markets Day
c	Coherence score
D	Document collection, corpus
d	Document
ϵ	Positive non-zero constant
EW	Entropy-based Term Weighting
θ	Document topic distribution
K	Number of topics
LDA	Latent Dirichlet Allocation
m	Confirmation measure
M	Set of confirmation measures
NER	Named Entitry Recognition
NLP	Natural Language Processing
NLTK	Natural Langauge Toolkit
NPMI	Normalized Pointwise Mutual Information
P	Word probabilities
PMI	Pointwise Mutual Information
S	Word subset
tf	Text frequency
TWLDA	Term Weighting Latent Dirichlet Allocation
V	Vocabulary
φ	Word probability distribution
w	Word
z	Topic

1. INTRODUCTION

The background of the present research is digital consultancy work selling technology, design, and strategy & culture solutions to a variety of companies in several industries. In consultancy work, having a consistent flow of clients and client projects is a key requirement for success, as the business model relies on work provided to clients. This dependency between obtaining and retaining clients and business success makes the role of sales crucial. In this research context, the companies acquiring consulting services are from different sectors ranging from digital service providers to the manufacturing industry, energy providers, and the health industry. Sales in digital, highly-specialized industries require background knowledge of the company and business sector from the salespeople [1]. To be able to tailor the offer, a sufficient amount of knowledge of the customers' business sector and current business situation is required. However, due to a large number of industries and companies in question, the amount of background work allocated for a single company must remain reasonably low to keep the cost of sales at a sustainable level. Hence, a tool to provide meaningful and easily interpretable insights into a candidate company's business, development goals, and aspirations would be beneficial for the sales process.

The present work aims to assist the work of salespeople by providing a semi-automated tool that generates insights into topics the companies of interest may be actively involved with, as based on the text materials they are publishing to the general public. As the companies regularly publish documents, reports, advertisements, and other material, the vast amount of information on available companies can cause pandemonium, halting the focus of the sales process if only a manual comprehensive analysis of the material is attempted. Current work aims to help with the sales initiation process and conversation-starting with the identified potential customer by providing an automated analysis of the current talking points of a company based on their publications. In addition, similar insights and reference information is extracted from the respective industry sector based on publications from several companies in the sector.

Applying data analysis and machine learning methods to support the sales process is not a new idea. However, the usual focus of such methods has been the lead acquisition and qualifying. Current studies provide data-enabled solutions for customer acquisition and lead qualifying, such as in [2]–[4]. However, only a few similar data-intensive solutions

exist for lead conversion and management [2]. This thesis aims to determine if similar data based solutions could be used to support the lead conversion by providing data-supported talking points for customer approach in a highly competitive environment.

1.1 Research Objectives

The objective of this thesis is to research whether it is possible to provide insights to improve the customer approach phase of the sales process. The research work tests the suitability of n-gram search and topic modeling as a method to gather actionable and understandable talking points from data published by companies. Discrepancies and abnormalities are discovered that separate and differentiate a chosen company from prior expectations of the sales experts and other companies in the same sector. The insights obtained from the modeling process are further used to spark conversation with potential customers.

The chosen research question for this research work is:
How to support the conversation-starting during the customer approach phase by analyzing material published by companies?

The study utilizes two separate datasets. *Company dataset* contains documents directly produced by an individual company and *Sector dataset* is a collection of documents created by companies within a distinctive sector. All the data utilized is publicly available without restrictions. The company and sector selection process is done by experts case-by-case. This thesis explores the suitability of topic modeling as a text mining method, in this case, Latent Dirichlet Allocation, to answer the research question. LDA is chosen as a method after initial research, as it is a widely used topic modeling approach and has proven to be successful in modeling a variety of datasets. The modeling results are analyzed with both qualitative and quantitative metrics. Finally, the suitability of the chosen approach is determined by combining the analysis of resulting metrics and expert opinions.

1.2 Structure of the Study

The work proceeds as follows. Section 2 introduces the theoretical background of the approach applied. The section introduces the technical approach by describing the topic modeling approach and evaluation metrics. A brief introduction on the sales process being improved is provided as well as a demonstration of the prior usage of topic modeling in the field of business analytics and sales support. Following, Section 3 defines the research approach and materials. The obtained data set is introduced, as well as the required preprocessing, and a description of the empirical setup for the study is presented. Section 4 reports the empirical findings and discusses the results. The evaluation of the

findings is described, thereby detailing the relevance and usability of our conclusions. The challenges and limitations of this study are described in Section 5. Afterward, Section 6 concludes this thesis work by describing the overall aims, results, and effects of the work. Finally, possible future development needs and approaches are discussed in Section 7.

2. THEORETICAL BACKGROUND

Here the relevant background for the thesis is introduced and the necessary scientific background for understanding the research fundamentals is presented. The chapter details and defines the sales process in consultancy, explains the relevant parts of Natural Language Processing (NLP) and topic modeling, as well as, presents topic modeling evaluation metrics to assess the results.

The chapter begins with an explanation of the sales process to provide a holistic understanding of the background this work takes place. The sales process is presented from start to finish and the customer approach phase is discussed in detail, as the focus of this work is in that phase of the process. This is followed with an introduction to topic modeling, which is the main analysis tool. After a general introduction to the topic, a detailed explanation of the topic modeling algorithm Latent Dirichlet Allocation (LDA) is shown. Finally at the end of the chapter, different evaluation metrics are presented and discussed to provide a background understanding for the result analysis and discussion.

2.1 Business-to-Business Sales in Digital Consultancy

Business to Business (B2B) sales refers to a sales process and business related transactions between two business entities, commonly companies. The sales process is traditionally linked with selling products, but here the focus is on the process of selling services and expertise in the form of consulting. Consulting services are here defined as a professional practice providing expert advice within a particular field. The services are here referred to as projects, often with a set time-frame and budgetw. The projects sold can be described as products which are primarily marketed to external customers, distinguishing them from internal work and representing the primary source of revenue to a company [5], [6]. As defined prior, the successful, consistent sales process is a key factor in a modern consultancy company, where the business is built on experts solving problems of customers instead of offering and selling a specific service or a product. However, building a sales pipeline that offers consistent work-flow and income is not an easy task in the hectic consulting environment. The pipeline must provide high utilization rates and consistent deliverables to customers in order to maintain profitability. Simultaneously, the work must be balanced internally within the organization to ensure satisfactory delivery

to the customer and a balanced internal resource utilization, such as billable hours per working consultant. [5]

Selling services in industrial, technology-driven business-to-business environment is required due to high level of specialization, making a sufficient process and the role and skills of sales professionals crucial [1]. The customer companies in B2B market operate and offer often big and complex solutions for their business cases. Similarly, the solutions required and expected from consultants are often large and complex. The services and products sold to highly specialized companies are tailored and customized for specific use cases and needs. To be able to specify a solution, a sufficient amount of knowledge of the company, its operations, and market segment is required.

The business agreements and deals done in B2B sales are often larger financial decisions, commonly requiring approval from multiple people [1]. The financial investment of a company may be large, and the sales person must be able to assure the buyers they are making the right choice. A good sales person also knows a sale might not happen in an instant when the product or a service is presented, but require multiple discussions and negotiations with different stakeholders. Hence, the aim is to give a sales presentation with a long-lasting impression in order to ensure the potential customer sees the supplier as a value-adding partner with state-of-the-art technology; someone worth doing business with and someone able to provide business value.

The aim of the sold products or services is to create net-positive assets to the buyer over certain period of time, i.e. to create customer value [6]. The offered service must benefit the customer, which makes customer value one of the key concepts in B2B markets. This is why sales, even in technology driven industries, is not about the technology or features behind the solution. The sales focus is on the ways the solution is able to solve the problems of the customer or provide ways to stand out in a market and create measurable business value. Products and services sold in B2B world most often bring financial profits by allowing customer to gain a larger market share, increase their own sales, or sell with higher profit margins. The benefits may also come from reduced production costs or service development. Similarly, the solution sold can provide new business potential by enabling the customer to offer new products or services, or improve existing features, thus providing higher quality services or products. Other options focus on developing and optimizing the processes allowing the customer to achieve faster lead times and more sales. However, to able to provide these specialized solution to the customer, the idea and the benefits need to be clearly communicated to the customer. Cohesive communication ensures a clear picture on how the solutions will enable the customer to reach more potential and quantify the additional value. [1]

A distinctive feature of business-to-business markets is the sparsity of companies [1]. The number of companies operating in same market segment offering similar products is often

relatively limited in B2B markets. This means a market or industry is often dominated by only a handful of companies. Even with a small number of companies the competition within a field can be hard making it important for each individual company to stand out from the rest and gain advantage compared to others.

2.1.1 Sales Funnel

The sales funnel is a way to conceptualize the phases during the customer acquisition process from potential opportunities to confirmed agreements [5]. The funnel illustrates the sequential narrowing of potential customers from a large selection of unqualified leads (suspects) to the small selection of closed customers. Managing each part of the funnel well is hugely important for the success of the process. The funnel is visualized in Figure 2.1, where the triangular shape refers to the number of potential customer available within each state. The process follows both linear and circular manner: Traditionally the potential customers move from one phase to the next one if they meet a set criteria. However, a potential customer can also be moved to previous phases and included for example to an other sales process.

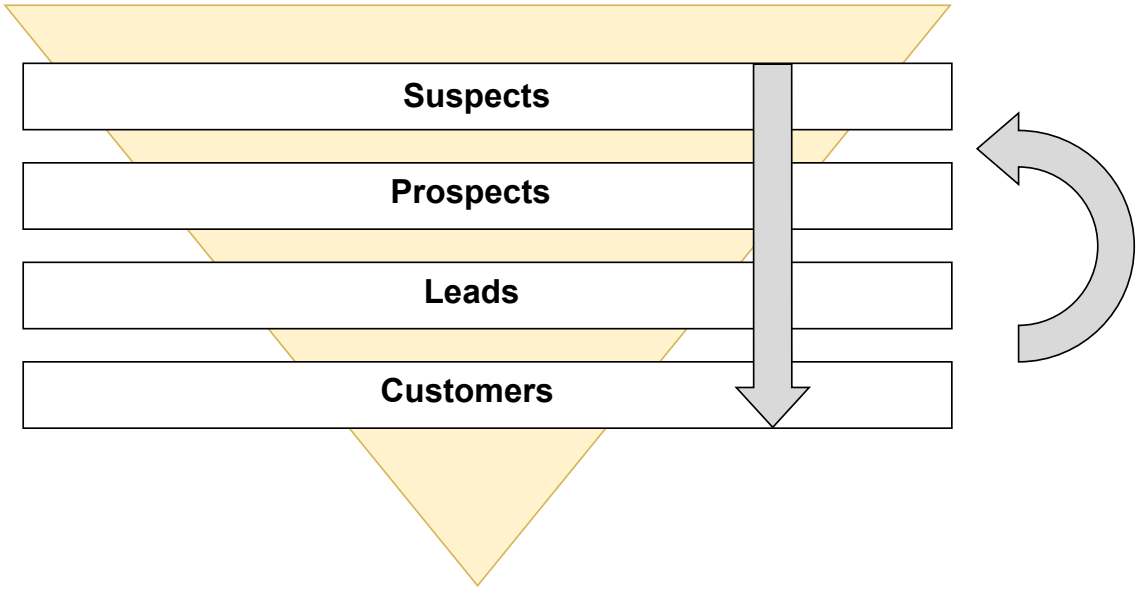


Figure 2.1. *Sales funnel. Adapted from [3]*

The funnel begins with a list of *suspects*, representing all available customers. In theory the list of suspects can consist of every company in B2B markets, but in reality the list is commonly limited to a selected companies within a specific industry [3]. To move forward in the funnel and in the sales process, a sales professional must make decisions on which companies to advance with. The decisions are based on the knowledge of the sales professionals and may be supported via additional analytics process. However, the vast amount of information available can be overwhelming, leading to decision-making based on arbitrary criteria and heuristics [3], [7].

Suspects which meet the predefined criteria, such as geographical location, industry, or company size, move to the next phase of the sales funnel and are referred as *prospects*. The next step is to further reduce the the number of companies as prospects to acquire *leads*, which are defined as companies that are likely to respond when contacted [1]. This phase is commonly assisted with data analysis to offer background information on the companies. After the lead qualification a selection of potential customers is approached. The approach is one of the most crucial phases during the process, as it is often the first interaction between the involved parties and will determine the continuation of the sales process. The aim of the initial approach is to begin a conversation between both sides. The interest of this thesis work is especially in the customer approach phase, and the conducted research aims to assist this step with additional insights. The customer approach is introduced in more detail in Section 2.1.2. Finally, with skilled sales people and a good offering, the leads become *customers* of the company after agreements between parties are made.

One of the most crucial decisions during the sales process is to choose which customers show enough potential to become prospects. Before approaching any potential customer, the prospect, referred as leads, need to be identified and qualified. These leads are generated, for example, from inbound sales from public presence, such as seminars, blog posts, etc., different data sources such as trade associations or market research companies, or by getting a reference from a current business partner. These leads are qualified based on varying criteria to determine the likelihood of a successful business co-operation, including the willingness and ability to purchase the offered product prior to approaching them. [1], [8]

Sales representatives are often presented with an overwhelming amount of information related to the company in question to support the decision making. The decisions need to be made in a fast-paced environment, yet be accurate to ensure the efforts are directed toward the most potential customers. More targeted and cohesive data helps the sales people make more accurate decisions and approaches leading ultimately to more successful leads and sales. [3]

Once the prospects have been analyzed and a point of contact has been decided a salesperson will approach the company, commonly via email or phone. It is important to try to reach the right personnel from the company to ensure the sales process continues. Similarly, it is important to create a relationship with the customer from the start by identifying mutual interests via small talk and adjusting the ways of being of the customer. The process can be supported by background information gathering prior to calling. This step is in the core of this research work and further discussed in Section 2.1.2. The approach should lead to a meeting with the customer. During this phase the salesperson will present the offer in more detail, but also carefully listen the customer to understand their business better. The salesperson aims to identify and describe the customer perceived

value and show it to the customer in convincing manner, and alleviate any concerns or questions the customer may have regarding the offer.

Finally, the process moves to the closing phase, where aspects of the deal, such as price and contract terms, are agreed and the focus will shift to practicalities of the offer. After the closing phase the agreed contract is put to practice and the development of the customer-supplier relationship begins. Well built and maintained relationships ensures future revenue and profits to both parties. Here the supplier must to ensure the conditions of the contract are met and the customer remains satisfied. Later on, new contracts can be made while the relationship develops and even long-term annual contracts or frame agreements can be agreed upon with successful customer relationship management. [1]

2.1.2 Customer Approach

After successful lead qualification the sales process moves to the pre-approach and approach phases. These two stages are closely intertwined and partially overlapping with each other. Approach phase defines the step when the customer in contacted for the first time making it especially important for the sales process. [9]

During the pre-approach and approach phases the lead is nurtured. The pre-approach phase focuses on acquiring more information on the client. This information includes, for example, the needs, habits, and preferences of a potential client. The information is then used in the initial contact with the customer by creating targeted and customized content for the potential customer as a way to nurture the lead. [9]

The approach focuses on building and establishing a trusting relationship with the customer, while acquiring more information to determine whether and how the offer could assist the customer. During the customer approach phase the first objective is always to reach the right people and spark a conversation. The aim at the beginning is not to sell, but to build a relationship and a conversation. Similarly to the pre-approach phase, the approach can include personalized and targeted messages to the potential customer, but also targeted content curation to make the company more visible.

In many cases, the first call or visit does not lead to a sale and more persistence is required to close the lead. Calling the customer without a pre-arranged appointment is called cold-calling. The leads for cold-calling are generated mainly by the potential customer approaching the company for example by visiting their website. The biggest challenge with cold-calling is often to find and reach the right person and, especially with bigger corporations. Reaching the right person in a position to drive the sales process. This may take multiple attempts and calls. Cold-calling as a sales method is generally not the most effective and most cold-calls lead to rejection. Better results are obtained from warm leads, in which case the person has has already showed some interest toward

the company and its services, for example, in a form of a website visit or other form of contact. These leads are more likely to spark interesting discussions. [1]

2.2 Natural Language Processing

Natural Language Processing (NLP) is an umbrella term for all machine-based processing of natural languages. It is a computational intelligence in the intersection of machine learning and linguistics. Natural language processing in its core aims to provide tools to process, analyze, and generate natural languages with a machine. It is a section of artificial intelligence that focuses on the processing and understanding of natural languages. The topics in the field of NLP include for example text mining, semantic analysis, machine translation, and text generation. These methods enable computers to analyze, transform, and generate natural language materials [10].

2.2.1 Topic Modeling

Topic modeling is a form of discrete data representation that provides an approach to finding hidden structures, also known as semantics, in large collections of textual data [10]. The method is commonly used for data mining, latent data discovery, and finding relationships within data in text documents. It is one of the most powerful techniques and widely applied on text set research on various fields ranging from sustainability analysis to medical research and other [11], [12]. An advantage of topic modeling is its ability to provide an automated procedure for breaking down large corpora and discover underlying themes from a corpus in a simple manner [13].

Topic modeling creates topics, collections of co-occurring words, and provides insights to the the collection as whole and to the relationships between documents. Topic models do not understand the context or concept of words in documents but find the words that best describe the information in the collection based on a statistical probability distribution created during analysis. A topic is defined as a set of words whose local (e.g., document-level) occurrence frequencies follow a particular distribution that is distinct from overall word frequency distribution across all documents of interest [10]. For instance, weather related terms, such as 'cloud', 'rainy', and 'sunny', should all be grouped together to define a weather related topic. The Figure 2.2 visualizes the idea behind topic modeling and shows the process from creating the word distributions to topic proportions and finally obtaining cohesive topics, human understandable topics. The process in Figure 2.2 is illustrative and does not represent real data.

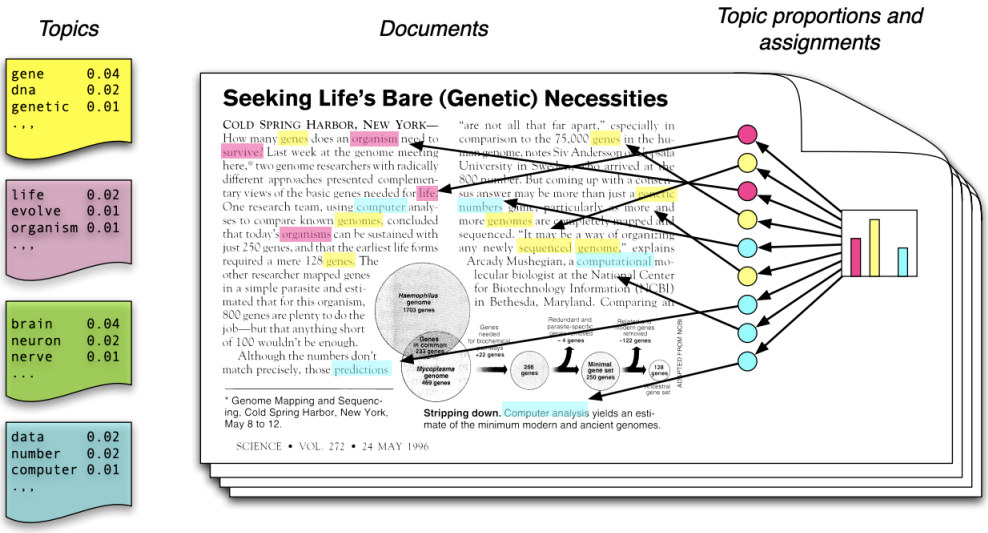


Figure 2.2. *Illustration of the idea behind topic modeling. Each colour represents a different topic. The method assumes a set of topics, which can be defined as distributions over a set words, exist for the whole collection and hence can be discovered. From [14]*

The topics obtained by the method have a varying degree of human-interpretability [15]. As the topics are learned directly from the collection of documents, some topics may incorporate words that have no clear mutual connection when analyzed by a human. One metric to control the topic quality and granularity is the number of topics. Topic modeling algorithms allow a strict control via topic number definition providing control over the granularity of the analysis [13]. The process of determining the optimal number of topics for desired granularity and interpretability via a coherence score analysis is further introduced in Section 2.2.3. Observing the topic coherence via a coherence score is a method to analyse the human interpretablity of a topic. The coherence score calculation and analysis process is further introduced in Section 2.2.3.

2.2.2 Latent Dirichlet Allocation

The conventional method for topic modeling is latent Dirichlet Allocation, LDA, which is a generative probabilistic model for collections of discrete data [16]. The algorithm was first introduced by Blei, Ng and Jordan in 2003 [16]. In general, LDA is a statistical method offering a powerful framework to represent and summarize contents of large document collections [17].

The method is able to create interpretable topics in an unsupervised manner requiring only little input and supervision from the user by inducing sets of associated words from the source text [18]. This simplicity of usage has allowed it to gain popularity. LDA is commonly used for text processing solutions, but the algorithm can also be applied to

other problems consisting of other forms of data, such as content-based image retrieval [14].

The underlying idea of LDA algorithm is to represent documents over latent topics. The model is based on an assumption that each collection of documents have latent topic in a from of multinomial distribution of words. Due to this statistical attribute, the top-N words with highest likelihood can be discovered [15]. The functionality of LDA is based on drawing upon word frequency and grouping documents with similar content into clusters [14]. The obtained clusters are referred to as topics. The LDA algorithm aims to estimate the posterior distribution of topics and topic proportions [13]. LDA is a form of a hierarchical Bayesian model utilizing three different levels to generate topics: topic distributions, topics, and words in a document. Each word is modeled based on an underlying set of latent topics and each topic is modeled based on an underlying set of topic probabilities. The topic probabilities provide a representation of the contents of a document or set of documents. [14], [16]

In LDA, a document is assumed to be generated by a generative statistical process, where each document contains similar statistical attributes. Hence the documents follow the principles determined below [14], [16]:

- *Document-topic relationship* determines the relative distribution of topics appearing in a given document. The Dirichlet distribution is defined as $\theta_d \sim Dir(\alpha)$, where θ_d determines the relative proportion of topics within a document. A random variable $\theta_d \in R$ is drawn for each document d in a corpus $\mathbf{D}$.
- *Word frequency* determines the distribution of terms in a selected topic. For each topic k a random variable is chosen from $\beta_k \sim Dir(\mu)$ to define term distribution in a topic.
- *Topic-word relationship* describes the relation between a word in a document and a topic. A topic is drawn from a multinomial distribution $z_t \sim Mult(\theta_d)z$, where θ_d is a prior. For every word, a scaled word frequency is defined by $tf_t \sim Mult(\beta_{zt})$.

The algorithm generates the topics using the process described below. It is noted that this representation follows the same simplification the used sources have followed. First, the dimensionality of the Dirichlet distribution, k, is assumed to be known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β, which is also treated as fixed quantity. Finally, the Poisson assumption is not critical and the document length distributions can be used when needed. Furthermore, it is noted that N is treated as independent from all other variables. [16]

LDA follows the detailed process for each document $\mathbf{w}$ in a corpus D [16]:

1. Choose $N \sim Poisson(\xi)$
   The Poisson distribution defines the likelihood of event occurrence over a time pe-

riod. The obtained value of N is independent from all other variables.

2. Choose $\theta \sim Dir(\alpha)$
   The Dirichlet distribution has a dimensionality of k that is assumed to be fixed for simplicity. The parameter θ represents a joint distribution of a mixture of topics.
3. For each word in $\mathbf{w}$ in document d:
   (a) Chose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$ representing a multinomial probability of words within the topic z_n.

The parameter β parameterizes the word probabilities as a $k \times V$ matrix, where V is the length of the vocabulary and k is the dimensionality of the Dirichlet distribution. The matrix represents the probabilities as $\beta_{ij} = p(w^j = 1|z^i = 1)$. [16]

Latent Dirichlet Allocation algorithm takes into account all three layers of the used data, corpus level, document level and word level, via the parameters it obtains. Corpus level parameters are α and β are selected at the beginning of the process and assumed to stay unchanged. Topic distribution θ is sampled on each document representing the document level and the parameters $\mathbf{w}$ and $\mathbf{z}$ representing the words in a document and topics, respectively, are sampled on every individual word in each document.

The Figure 2.3 provides a visual explanation of the algorithm. The picture defines the process of topic z selection for a word w. The outer square represents the number of documents M, that are used to define the Dirichlet parameter θ, which is dependant on the contents of the corpus in question. The inner square notes the number of words in a document N. The parameters α and β are defined only once and are not dependant on the modeling process. The process is repeated to all words and documents within a corpus. [16]

Mathematically, the process is defined as follows [16]. Given the known parameters α, β and N defining the length of sets, the definition of the joint distribution of topic mixture θ, set of topics $\mathbf{z}$ with length of N and set of N words $\mathbf{w}$ is defined as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{2.1}$$

where $p(z_n|\theta)$ is the topic distribution θ_i and the probability density $p(\theta|\alpha)$ of the k-dimensional Dirichlet variable θ is defined as:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}, \tag{2.2}$$

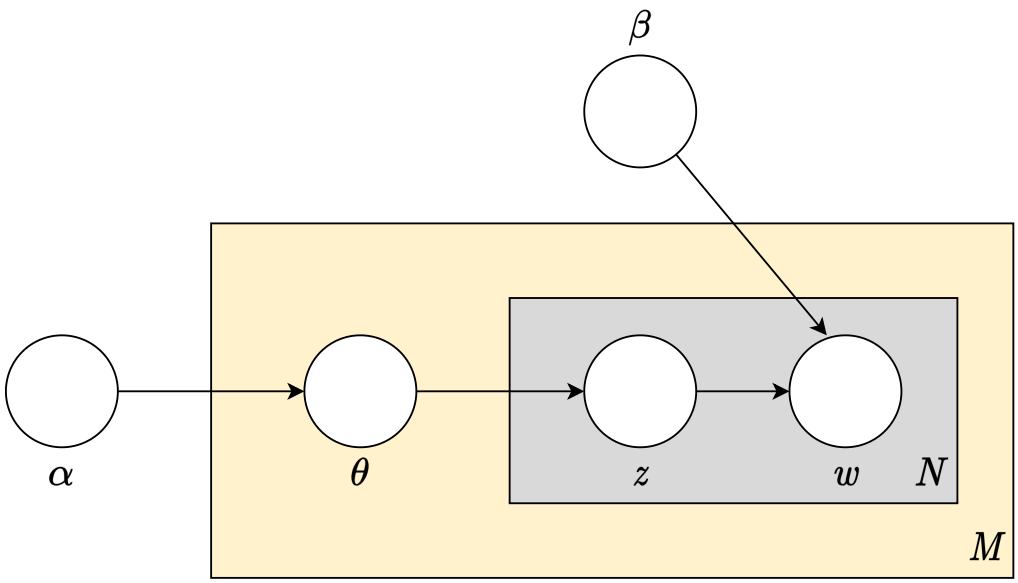


Figure 2.3. *: Graphical model representation of LDA. Adapted from [16]*

where α is the k-vector with values larger or equal to 0 and $\Gamma(x)$ is a Gamma function.

The marginal distribution of a single document is obtained by integrating over the value θ and sum over the topics z:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta))d\theta, \tag{2.3}$$

where $\mathbf{w}$ defines the sequence of words within a document.

Finally to obtain the probability of the whole corpus D, the product of the marginal probabilities of each document is calculated:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{d,n}} p(z_{d,n}|\theta_d)p(w_{d,n}|z_{d,n},\beta))d\theta_d. \tag{2.4}$$

The Equation 2.4 shows the probability of selected corpus given the parameters α and β.[16]

2.2.2.1 Parameter Estimation

In practice the parameters used in LDA are calculates from the source data during the learning process. The parameters and distributions are updated by iterating over the source material, which makes topic modeling a computationally heavy method [19]. For example Laplace approximation, variational approximation, and Markov chain Monte Carlo can also be used as approximate inference algorithms for LDA, but here the Bayes method

is used. [16].

The parameters α and β are found from the source text by maximizing the log likelihood of the data, marginalizing over the hidden variables θ and z:

$$l(\alpha, \beta) = \sum_{d=1}^{M} logp(w_d|\alpha, \beta) \tag{2.5}$$

The calculation is done using empirical Bayes estimates via an variational EM procedure consisting of 2 steps [16].

1. **E-step**: Compute the optimizing values of the variational parameters for each document D. In practice the values computer are:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta}{p(w|\alpha, \beta)} \tag{2.6}$$

2. **M-step**: Estimate the parameters α and β by maximizing the lower bound on the log-likelihood with respect to parameters.

The two steps are repeated until the lower bound on the log likelihood converges [16].

2.2.2.2 Discovering the Optimal Number of Topics

The selection of appropriate number of topics, denoted as K, is a key metric to a successful application of topic modeling algorithms [20]. Too low value for K causes the model to become too undefined to produce coherent topics. The topics become overly generalized and contain a wide range of terms that are hard to interpret under a cohesive theme. Correspondingly, a too large value of topics leads to overly complex model. The quality of results deteriorates due to over-clustering, where the corpus is sectioned to specialized but highly-similar topics that are hard to interpret.

Three variables are be selected prior to modeling: hyper-parameters α and β, and the number of topics K [21]. All the variables have a substantial importance to the modeling as the parameters affect the dimensions and a priori defined distribution of target variables ϕ and θ. The variables heavily affect the quality of the models and the resulting topics. The α and β parameters are introduced in prior Section 2.2.2. In this section the focus is on introducing the characteristics of defining optimal number of topics.

The selection of the number of topics is not trivial. Coherent topics may be obtained from a corpus with different values of K, as topics can be present at several resolutions from fine-grained to coarse [20]. Hence, one corpus may contain multiple suitable values for K depending on corpus structure and end-use of the analysis outcomes.

The literature suggests there is currently no standard method for the procedure of defin-

ing the best performing number of topics. A common method is to run several candidate models with different values for K (e.g. in [13], [22]). The resulting models are compared for differences and interpretability [21]. However, the method described has no quantitative metric to measure the suitability of the values of K. Due to this, some researchers suggest the usage of further criteria. One commonly used metric is *perplexity*, which is further introduced in Section 2.2.3.1 (e.g. [16]). Perplexity is a metric that defines how well a model predicts a held-out section of the corpus and provides a comparable metric between different models [21].

2.2.3 Topic Model Evaluation Metrics

Different evaluation metrics that are used to assess the performance of a trained model. The section discusses the different methods on topic model evaluation and introduces the relevant metrics.

The measures introduced here are based on statistical methods. The metrics produced cannot provide an absolute score defining the goodness of a model, making human ranking still the most reliable way to assess the performance of a model. However, human assessment is yet expensive to produce, favouring the usage of mathematical scores [23].

2.2.3.1 Perplexity

Perplexity is a statistical metric defining the prediction accuracy of a trained model when introduced to new data. Perplexity is a metric calculated as the inverse of the geometric mean per-word likelihood [16], [24]. Mathematically the metric is defined as [16]:

$$perplexity(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d}\}, \tag{2.7}$$

where D refers to a test document in a test set of M documents and w is a word in a document of N words.

The metric is especially useful for model selection among alternative generative models and for demonstrating a performance advantage of a model compared to another one. Perplexity measures are also often used for model parameter tuning, such as determining the optimal number of topics, as it provides an easy way to compare model performance [16], [24]. In general, a lower value for perplexity indicates better performance and a higher ability to generalize between datasets [16]. The full details of the calculations are presented in [25], where the authors presents an efficient and unbiased computing approach for perplexity calculations.

However, the metric has also received criticism. Chang, et al. (2009) in [26] presented

the first ever human-evaluation of topic models, in which humans were asked to identify discrepancies in topics. The work showed that humans in some cases preferred models with a higher perplexity, whereas lower perplexity has commonly been defined as better performance [26]. Due to the counter-intuitive results of the study, perplexity may not be an optional metric for model topic model evaluation.

2.2.3.2 Topic Coherence

The topics created by a model are not guaranteed to be interpretable. A metric commonly used to examine the performance of a model and to distinguish between well and poorly performing models is topic coherence [27]. The metric approximates the extent to which the topic modeling results are interpretable to humans, ranking the topics based on interpretability similar to human assessment [24]. Coherence score is calculated separately for each topic, where it measures the semantic similarity of the top words of each topic. Topic coherence is not an explicit performance metric, but can provide a general view to assess the models ability to construct cohesive topics [15].

A general framework for topic coherence calculation is presented in Figure 2.4. The workflow consists of four steps: Segmentation, probability calculation, confirmation measure calculation, and result aggregation. The final output is the coherence score c of a topic t. During the segmentation step the set of words t is segmented to a collection of word subsets S. Next the word probabilities P are calculated with a reference corpus. Both variables S and P are then passed to a confirmation measure to obtain an agreement φ of pairs of S. Finally, the values are aggregated to a single coherence value c. [23]

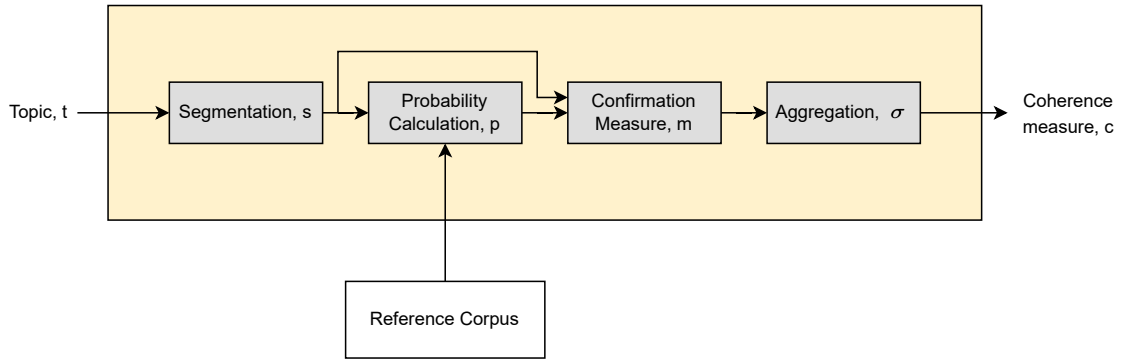


Figure 2.4. *General coherence calculation framework. Adapted from [23]*

In summary, the framework in Figure 2.4 calculates a cross-product of the four sets present as the coherence value c :

$$\mathcal{C} = \mathcal{S} \times \mathcal{M} \times \mathcal{P} \times \sum \tag{2.8}$$

In Equation 2.8 S is the set of segmentations, M defines the set of confirmation mea-

sures, P represents the word probabilities, and $\sum$ defines the used set of aggregation functions. [23]

Topic coherence is commonly calculated on the top N words of a topic with N ranging between 10–50, as they commonly determine sufficiently the subject of a topic. Newman, et al. in [24] introduce and test multiple topic coherence measures using two distinct datasets consisting of a collection of news articles and a collection of books. The coherence scores of each topic are calculated with the 10 top words of a topic. Their study defines the Pointwise Mutual Information (PMI) as the most consistent and best performing metric. The approach is able to reach top or near-top results on both datasets. Other metrics introduced in the study are based on, for example, cosine similarity, category overlap, and semantic similarity, but those methods are not proven as successful as the PMI score [24]. The PMI score has often the highest similarity compared to human ratings of a topic [23].

The Pointwise Mutual Information score is calculated by term co-occurrence via simple pointwise mutual information. The PMI score is computationally manageable and that the approach of focusing on the word-pair based co-occurrence is a highly successful method on coherence calculations. [24] The PMI score is calculated over all the word pairs (w_i, w_j) in a topic using a sliding window of 10 words, also known as a context vector [28]:

$$PMI(w_i, w_j) = log\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{2.9}$$

The standard Pointwise Mutual Information score is further improved by normalizing the elements of used context vectors and treating the vectors as normalized PMI ($NPMI$) [23]. A context vector for a word w is created using the counts of word co-occurrences using a context window which includes all words within 5 tokens of the selected word. The Normalized Pointwise Mutual Information, $NPMI$ (originally from [29]), for a pair of word (w_i, w_j) is defined as [23]:

$$NPMI(w_i, w_j)^\gamma = (\frac{log\frac{p(w_i,w_j)+\epsilon}{p(w_i)p(w_j)}}{-log(p(w_i, w_j) + \epsilon)})^\gamma, \tag{2.10}$$

where the γ parameter is a weight parameter where increasing the value gives more weight to higher $NPMI$ score values and ϵ is added to avoid division with zero [23]. The scores produced by the metric are in range of [-1, 1], where higher positive value refers to higher similarity of top words [30].

Another metric commonly used is UCI score. The UCI coherence metric utilizes the PMI score with additional scaling parameter based on the number of words N within a topic. The score is traditionally calculated with PMI but the metric performs better when the PMI score is replaced with normalised variant $NPMI$ [15], [23]. The UCI coherence

function is defined as [23]:

$$C_{UCI} = \frac{2}{N \times (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} NPMI(w_i, w_j) \tag{2.11}$$

The $UMass$ utilizes a different calculation metric. It uses an asymmetrical confirmation measure between top words as smoothed conditional probability [17], [23]. The score accounts the order of the top words within a topic as an additional feature accompanied by $NMPI$. The word probabilities are defined based on the document frequencies on the original set of documents used for the training. The score is calculated as [23]:

$$C_{UMass} = \frac{2}{N \times (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}, \tag{2.12}$$

where the ϵ parameter is added to the coherence to avoid potential $log(0)$ - calculation. The standard value in prior usage is $\epsilon = 1$ but the metrics perform better when ϵ is chosen to be close to zero [31].

It is worth to note that as calculated coherence metrics are commonly based on the number of word co-occurrences, the traditional coherence scores work well for long documents but are less favorable with short texts [17].

3. DATA AND METHODS

This chapter provides an overview of the research material and methodology, as well as a reasoning behind conducting this study. The used data sources and source data qualities are described in detail. The Methods section defines the why and how for the research and the following sections detail the implementations used. Finally, the section concludes with description of the experimental setup used and the evaluation criteria followed on assessing the results of this work.

3.1 Data

The source material utilized consists of different business related publications. All the source material is publicly available and published by companies. Companies publish a variety of documents to provide public disclosure on the effects of past decisions, their current advances, and future aspirations. The specific documents used in the corpora are capital market day (CMD) presentation visuals, transcribed capital market day speeches, job postings, and press releases. None of the documents are legally required from a company to publish, and hence follow a free-form formatting and contain different topics that are not put under external scrutiny to ensure, for example, accuracy. CMD materials and job postings are future-facing and incorporate future aspirations and plans, and press releases contain the latest advances the company wishes to make public. Documents such as annual reports and financial reports are ignored due to the historical focus as the aim is to discover future aspirations rather than examine past actions.

The documents are published for different stakeholder groups to create a honest but favourable image of the state of the business operations. The stakeholder groups considered with selected publications are investors, employees, prospective employees, customers, suppliers, communities, governments, and similar parties impacting the company. Depending on the document the focus groups vary. **Capital market day** - materials are mainly directed toward current and potential investors. The aim of the material is to provide a better and deeper understanding of the business operations and financial status of the company to the investors. The companies commonly introduce key priorities for upcoming years, future development aspirations, profitability metrics, and core focus points for long-term strategic approaches. The top leadership of the company commonly

Table 3.1. *Source material description*

Dataset	Num. of documents	Publication years	Avg. length (characters)
Company	108	2022	12181
Sector	375	2017 – 2021	27748

presents the metrics and topics to the invited investors and analysts. **Job postings** are aimed for potential employees to describe available vacancies, key benefits, and values of the company. Job postings define the current staffing needs and can determine both current operations and future development aspirations via talent acquisition needs. **Press releases** are published on selected, often on high-profile changes or updates focusing on the key aspects of the company. Press releases commonly define the highlights and the most important events giving insight to the values of the company. However, all the analyzed documents aim to create a favourable image of the company itself and its' operations. The tone of the document will likely be positive and the documents are produces by or in co-operation with marketing departments to ensure cohesive outlook of the company. Due to this, the results from the analysis results are likely to be relatively positive.

Two separate datasets were gathered: A company dataset containing documents of an individual company and a sector dataset containing documents from multiple companies on a chosen sector. The corporas used contain the material introduced in the previous section: Job postings, capital market day materials, and press releases of chosen companies. The case companies were chosen based on expert opinions as well as document availability and operating market. Table 3.1 presents the key details of the source material.

3.2 Methods

The research conducted focuses on finding how the sales process can be augmented trough the utilization of text-based analytics methods. In detail, the work focuses on discovering actionable insights to support the sales professionals on their approach and conversation-starting with potential customers. The aim was to gather concrete talking-points with the assumption that the materials published by company should indicate the most important topics to them, but also help to discover topics the company is not talking about.

The hypothesis was tested by building a pipeline consisting of preprocessing, n-gram search, and topic modeling. The Figure 3.1 visualizes the created framework for the research and steps within. Each step of the research framework is described in detail below.

During the development the performance of the topic modeling algorithm was evaluated

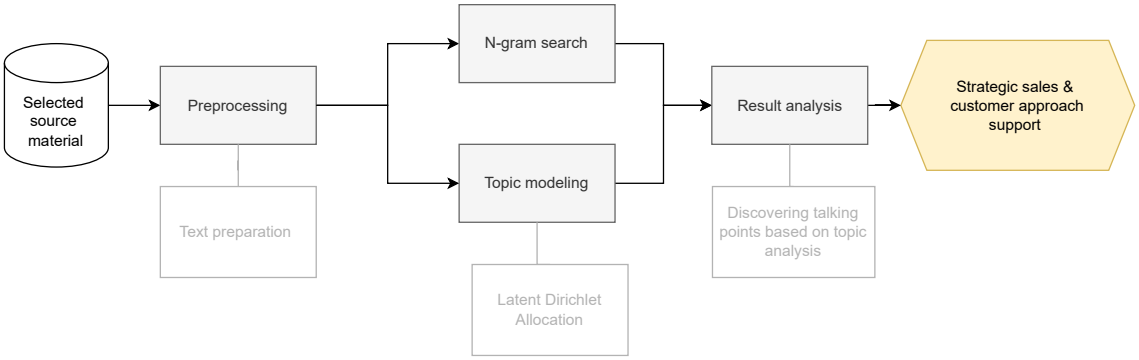


Figure 3.1. *Research framework visualization. Adopted from [13]*

in a quantitative manner using the coherence scores. This process ensured the most suitable parameter selection for the modeling process, but does not directly assess the quality of the final resutls. The final results are evaluated in a qualitative manner. Qualitative analyses of the results were conducted by presenting the system outputs to two expert assessors, who were asked to comment on the usability of the system and the quality of the results using free-form description for each of the outputs. The experts had multiple years of work experience in consulting work with one having a strong sales and leadership background and the other having a strong technical background with experience from the sales process. The assessment was done in a unstructured and conversational manner. The exact questions for the assessment are presented at the end of this Chapter in Section 3.2.3.

3.2.1 Data Preprocessing

The corpora utilizes real-world data and hence contains relatively large amounts of noise distorting the analysis results. Each company presents the data in a different structure. The research also utilizes transcribed speech material, which requires extensive preprocessing due to the incoherent nature of free-form human speech. The built pipeline includes multiple preprocessing steps to obtain a cohesive dataset for the analysis phase. The steps in the preprocessing pipeline are described below and appear in order of execution.

The gathered data is in PDF or textual format depending on the source. The textual data does not require any additional steps prior to preprocessing, but the PDF files must be converted to text for preprocessing. Prior to preprocessing the pipeline utilizes Google's Cloud Vision API[1] to convert PDF's to text. The Vision API works well for PDF documents that have visual elements, such as timelines and text blocks, but is relatively slow for a larger amount of documents and requires payment.

The documents gathered are mainly in English, with some exceptions and some multi-

[1] https://cloud.google.com/vision/docs

lingual documents. The language is checked using SpaCy[2] and any document not majorly in English is omitted during the preprocessing step for simplicity. After the language check, the preprocessing continues with **lowercasing** step, during which all letters are converted to lowercase. The process is simple and creates effective results, especially in cases the analysis requires recognition of terms or outputs. A machine differentiates between a lowercase letter and a uppercase letter, 'a' is different from 'A', but the semantic meaning in text analysis is commonly the same. Removing the capitalization results a better accuracy outputs and better coherence on analysis results. Next the pipeline performs a **noise removal** step. The texts contain noise in form of symbols, punctuation marks, numbers, white spaces, and other which are not necessary for analysis purposes. A better result accuracy is obtained after removing noise from the texts. The removed noise is case sensitive and must be adapted to the use case. The source material used contains a high number of stop-words, especially in the transcribed speech, as humans tend to use filler words in their speech. As a part of the noise removal, stop-words are removed from the text. The removed list of stop-words is combined from three different sources:

- a default English corpora stop-words from Natural Language Toolkit (NLTK) [3],
- a list of words that have been manually curated for the use-case, and
- a list of named entities, such as names, organizations, and dates, curated from the text by a Named-Entity Recognition (NER) search.

The manually curated list of stop-words is presented as an Appendix A.

As the final step of the preprocessing performs **lemmatization** to the text set. Lemmatization is a process of returning a word to its dictionary form and removing any inflectional endings [18]. This allows similar words to have the same information value regardless of conjugation of the word. Lemmatization is a similar process to stemming where the word is reduced words to its root form. However, in lemmatization the process of reducing to words is more sophisticated than in stemming and follows a process that abides by rules. The rules ensure the lemmatization process produces results with a higher accuracy. The advanced process uses commonly mapping and can reduce word forms, for example, from comparative or superlative forms.

A semi-automated preprocessing pipeline ensures easy usage of the system. Figure 3.2 visualizes the steps on the preprocessing pipeline. The pipeline is implemented using Python and build-in functionalities from NLTK [4].

[2] https://spacy.io/
[3] https://www.nltk.org/howto/corpus.html?highlight=stopwordsword-lists-and-lexicons
[4] https://www.nltk.org/

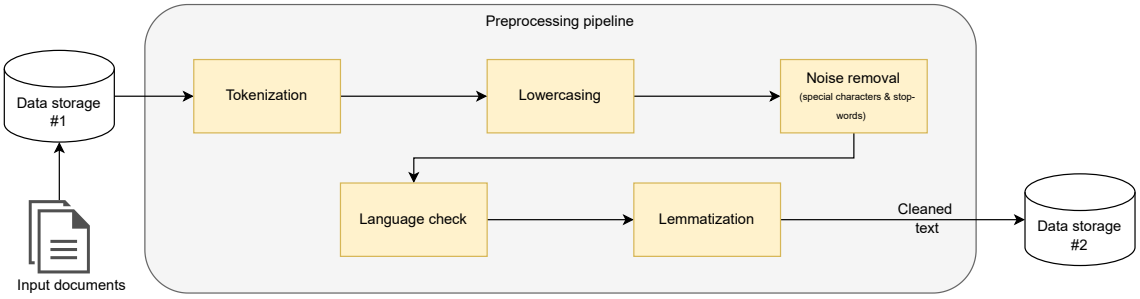


Figure 3.2. *Preprocessing pipeline*

After the pipeline finishes the results are stored for running the analysis using n-gram search and topic modeling. The details of the experimental setup are presented in Section 3.2.3.

3.2.2 N-gram Calculation and Topic Modeling

The N-gram calculations are implemented with Python and consisted of two separate steps. Firstly, the words are grouped together with the n following terms to create n-grams. The calculations allow for any value of n, but for this research values of n = 1,2,3 were used. After this, the amount of occurrences per term are calculated and the 10 most common unigrams, bigrams, and trigrams were presented for clarity. By default the n-grams are calculated for the whole dataset, but a filtering option allows to choose for example a specific year or source data type.

The research uses a topic modeling algorithm called Latent Dirichlet Allocation as the method of topic modeling. The topic modeling algorithm is run using Python and a LDA implementation from a open-source Python library `Gensim`[5] and required scientific Python libraries. The algorithm allows some fine-tuning to the used parameters. The default values provided by Gensim are following:

```
class gensim.models.ldamodel.LdaModel(
corpus=None, num_topics=100, id2word=None, distributed=False,
chunksize=2000, passes=1, update_every=1, alpha='symmetric', eta=None,
decay=0.5, offset=1.0, eval_every=10, iterations=50,
gamma_threshold=0.001, minimum_probability=0.01,
random_state=None, ns_conf=None, minimum_phi_value=0.01,
per_word_topics=False, callbacks=None, dtype=<class 'numpy.float32'>)
```

The Gensim documentation offers the details of the model usage [6].

[5] https://radimrehurek.com/gensim/
[6] https://radimrehurek.com/gensim/models/ldamodel.htmlgensim.models.ldamodel.LdaModel

3.2.3 Experimental Setup and Evaluation Criteria

The experimental setup is built using Notebooks on Google Colab using Python. The approach does not require local installation and the tool can be used on the browser. The preprocessing phase and the analysis phase are separated to different notebooks for clarity. The used data is stored in Google Drive for easy access and usability.

The experimental setup begins with a preprocessing phase concerning a number of prepossessing steps to obtain a clean dataset to ensure the best possible analysis results. The filtering rules applied are introduced in Section 3.2.1. After the preprocessing, the setup includes two lines of analysis: N-gram search and topic modeling. In the N-gram search, the most common uni-, bi- and trigrams of the dataset are calculated. The approach is based on counting the number of occurrences for each word or pairs of words. The method provides a list of the words most commonly present in a dataset.

Independently from the n-gram search a second line of research is conducted. A topic modeling algorithm, in this case Latent Dirichlet Allocation, is applied to the dataset. Prior to the application of the algorithm itself, some exploitative steps are performed to fine-tune the used parameters. The parameters tuned are the standard LDA variables number of topics k, α, and algorithm parameters 'chunksize' and 'passes'. The most suitable values for each parameter were defined using an iterative manual process and observing the development of coherence scores with different parameter combinations. After parameter definitions and modeling the source corpus, the topics are allocated for further analysis to support strategic decision making and sales approaches to potential partners.

After the analysis the system to obtain the results and results are evaluated. The evaluation criteria used is twofold: The usability of the system without extensive technical knowledge and the usability of the analysis results on supporting the sales process. Both of the assessments are done in a unstructured discussion with experts. The usability discussion assessment included the following themes:

- Can the system be used without the need to install anything locally?
- How is the data stored and how can the user add more data?
- How easy is the analysis code to run? Any variables or credentials that need to be updated?

The result analysis discussion included the following themes:

- How are the results presented? How easy are the results to understand without technical knowledge?
- Is there a need for further analysis prior to using the results during the sales process?

- Are the results specific enough to provide actionable insights?
- Are the results trustworthy and reliable enough to be used in during the sales process?

The results of this thesis work are assessed based on this criteria and the evaluation is presented with the results in Chapter 4.

4. RESULTS AND EVALUATION

The fifth chapter introduces the results of the conducted research. The first two sections focus on presenting the results of the search for the best performing combination of parameters with LDA. Following, the Section 4.3 presents the results of the n-gram search and topic modeling, and discussess the contents of the created topics in detail. Finally, the chapter is concluded with a evaluation of the results of this study.

4.1 Optimal Number of Topics and Parameter Tuning

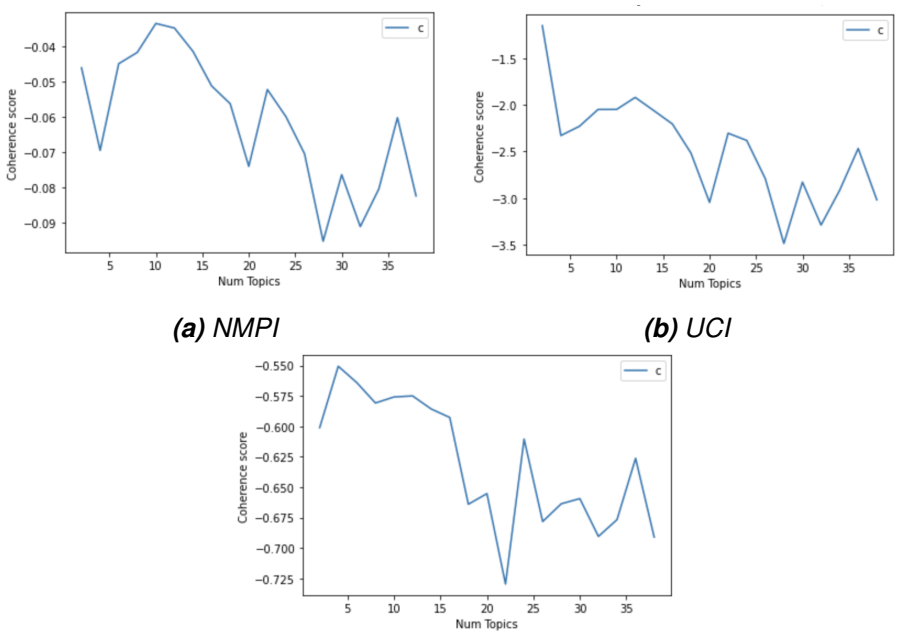
The quality of the modeling results and topic coherence depends highly on the hyperparameters of the LDA algorithm chosen prior modeling. The parameters commonly used are Dirichlet distribution variables α representing the document-topic density and β defining the topic-word density, and the number of topics K [16].

Number of topics K is one of the parameters that is defined during the model training and the coherence score is highly affected by the number of topics. The selection for the most suitable number of topics was done using an iterative algorithm to test the modeling coherence on different number of topics. The coherence scores were calculated with different metrics and different number of topics to find the overall most suitable number topics. The results of both datasets are presented in Figures 4.1 and 4.2 corresponding to the company and sector data, respectively.

Multiple coherence metrics were used during the calculations to get a better understanding and estimate of the effect to the model quality. The dataset used was relatively small and the values to test the coherence score development are from K = 2 – 50 for the sector data and K = 2 – 40 for the company data, with an increment of 2 after each run.

Figure 4.1 presents the evolution of coherence values for the company dataset and Figure 4.2 shows the coherence score value development with the sector dataset. The coherence scores are not directly comparable, as the values produced are on different scale for each coherence metric. Hence, the focus is on the evolution of the coherence rather than absolute values. The different coherence metrics are introduced in detail in Section 2.2.3.

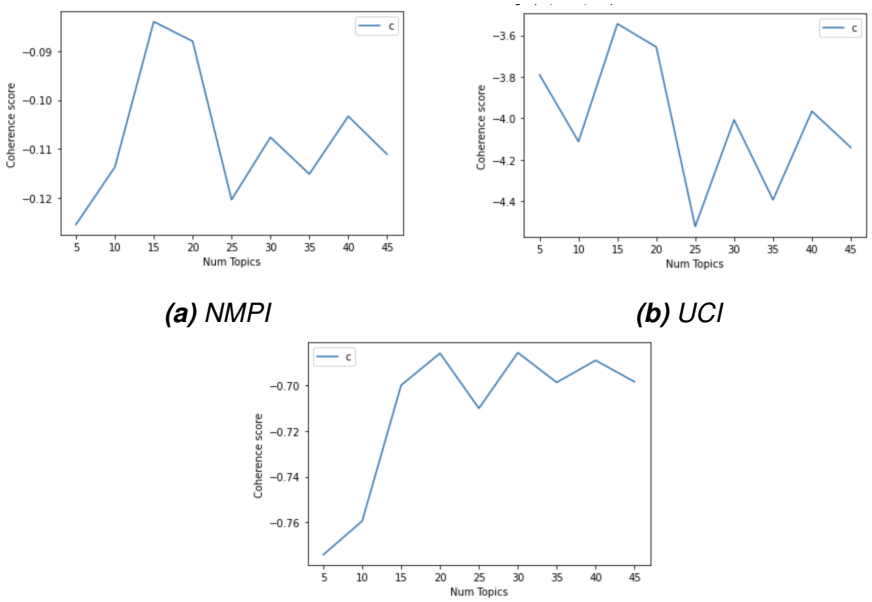
The graphs in Figure 4.1 show higher values on the lower values of K. The quick decline



(a) *NMPI* ***(b)*** *UCI*

(c) *UMass*

Figure 4.1. *Different coherence score values corresponding to different number of topics, company dataset*



(a) *NMPI* ***(b)*** *UCI*

(c) *UMass*

Figure 4.2. *Different coherence score values corresponding to different number of topics, sector dataset*

on the graphs' values around between 10 – 15 suggest the most suitable value for number of topics K is before this area. The graphs suggest the most suitable K would be around 10 and the final choice for the number of topics K = 9. Following, the graphs in Figures 4.2a, 4.2b, and 4.2c present similar findings on the coherence score development for the sector dataset. The graphs reach the highest values between the range of 15–20

Table 4.1. *Topic coherence values*

Coherence metric	Score / Company data	Score / Sector data
$NMPI$	-0.0872	-0.1112
C_{UMass}	-0.6531	-0.6930
C_{UCI}	-3.1457	-4.1380

and after that drop significantly. The graph 4.2c demonstrates the smallest drop and the values remain overall higher in the $K \leq 25$ range. Based on the graphs, the optimal number of topics is in the range of K = 15–20 for the sector dataset and the final choice for the use case is K = 15.

4.2 Coherence Evaluation

The resulting topics are evaluated using the coherence metrics introduced in Section 2.2.3. The Normalised Point-wise Mutual Information ($NMPI$) score is used as the base metric for the coherence analysis. Table 4.1 presents the values of used coherence calculations. The $NMPI$ values range is [-1,1], where higher positive value defines more cohesive topic. The score obtained for $NMPI$ is near zero, but on the negative side. This implies the topics generated by the model are not coherent and easily interpretable. However, the score is not highly negative implying some of the topics have reached a good interpretability, yet overall the model performance is mediocre.

To further prove this statement other coherence metrics are utilized: $UMass$ and UCI. The values of each metric are presented alongside $NMPI$ in Table 4.1. It is important to notice that due to different ways of calculation and ranges, the metrics are not directly comparable, but can provide insight on the general model performance. For the $UMass$ the optimal score is zero, which refers to mathematically perfect coherence. The negative score with clear distance to zero further supports the prior assessment of the model quality, as the score is based on how often two words of the top words appear together in corpus. Similarly, the negative UCI score supports the argument of incohesive topics.

4.3 Topic Modeling and N-gram Search Results

The n-gram search results were the first obtained and the n-grams gathered from company and sector dataset are presented here. The figures presented show the unigrams, bigrams, and trigrams calculated from the dataset, respectively. Figures 4.3, 4.4, 4.5 display the top 10 most common n-grams present and number of total occurrences per n-gram in the company dataset. Similarly, Figures 4.6, 4.7, 4.8 present the top 10 most

common n-grams with number of total occurrences in the sector dataset. The personnel and company names present in the both datasets and results are disguised from the results for privacy reasons.

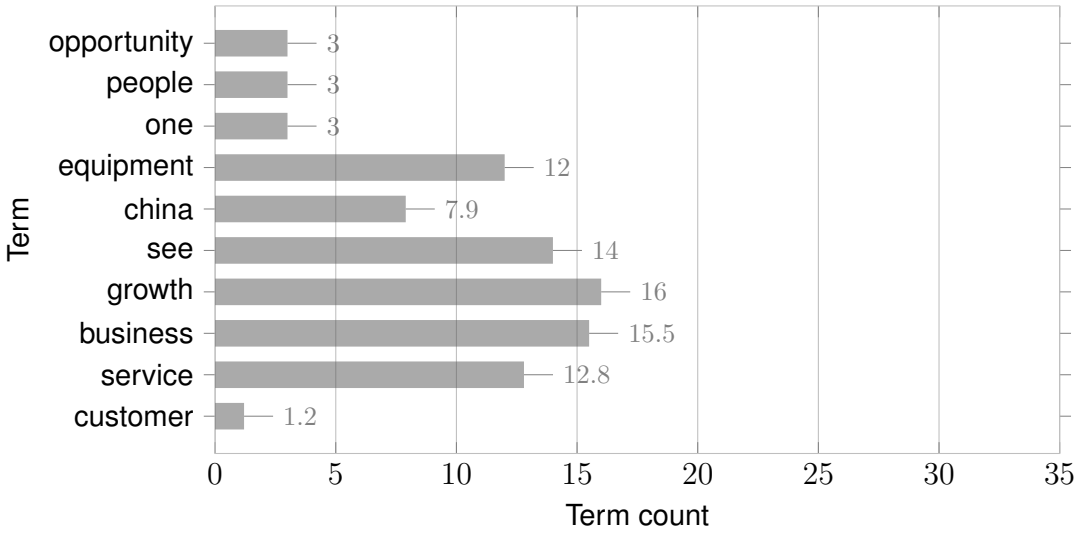


Figure 4.3. *Top 10 most common unigrams, company dataset*

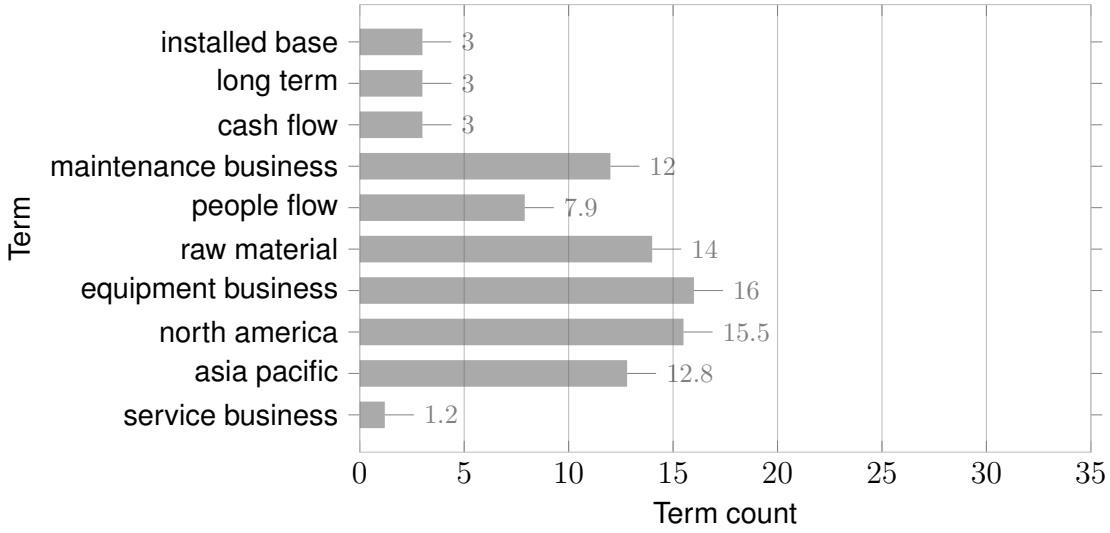


Figure 4.4. *Top 10 most common bigrams, company dataset*

The second part of the results focuses on the topics acquired from the modeling process. Latent Dirichlet Allocation (LDA) was the used method of topic modeling. The algorithm is introduced in detail in Section 2.2.2. The topic modeling results are presented in Tables 4.2 and Table 4.3 for an individual company and for a sector, respectively. Each topic collection contains the top 10 words of the topic and the words are presented in order of likelihood within each created topic. For the company dataset, all modeled topics are presented. For the sector dataset, a selection is presented, as the number of topic used is higher than for the company dataset. The full sector dataset modeling results are presented in the Appendix B.

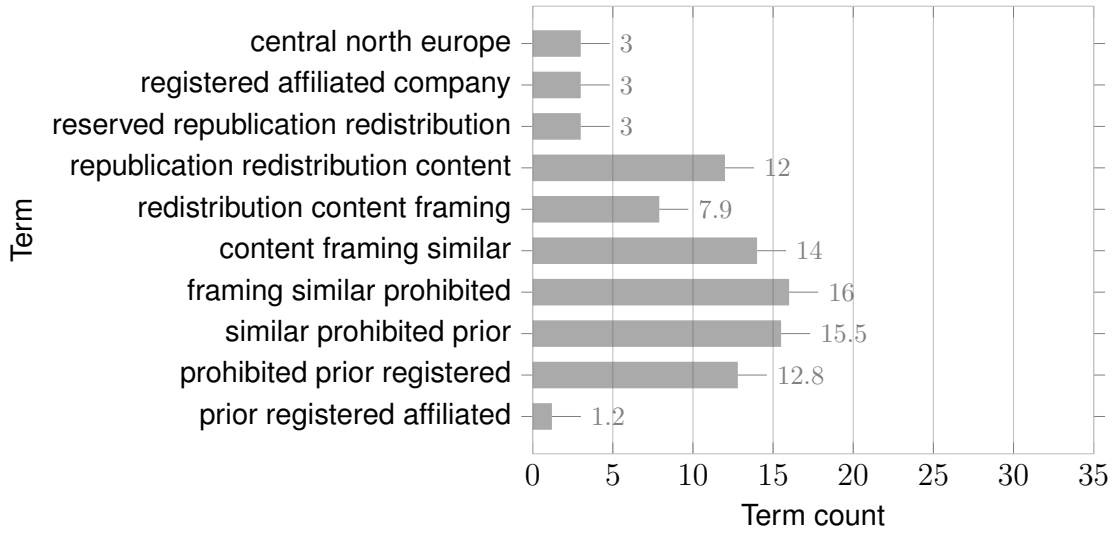


Figure 4.5. *Top 10 most common trigrams, company dataset*

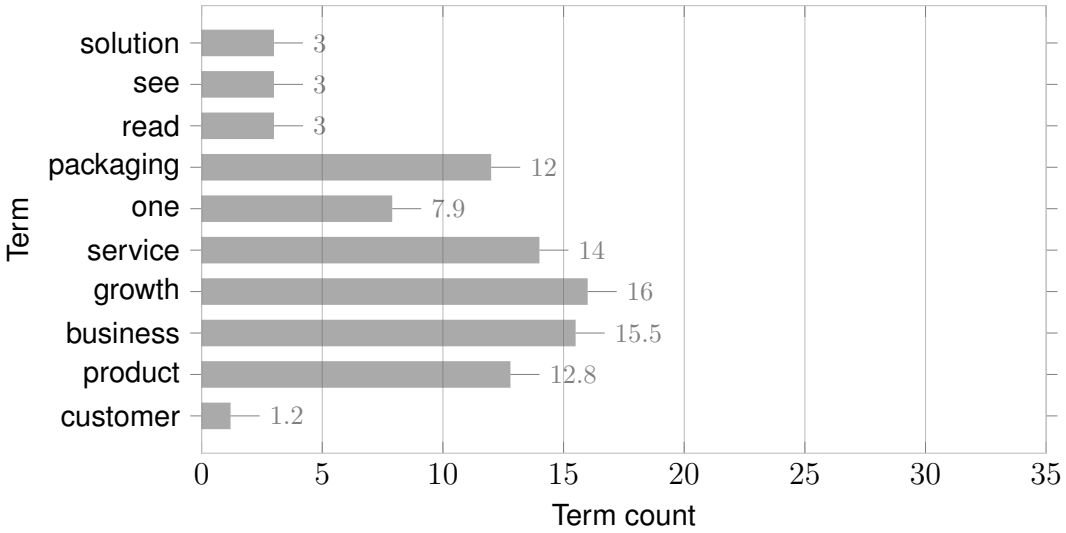


Figure 4.6. *Top 10 most common unigrams, sector dataset*

The interpretability of the company related topics presented in Table 4.2 fluctuates, but overall the LDA algorithm generates a few coherent topics. Additionally, in a few topics the modeling process or the preprocessing sequence has caused the beginning of a word to be missing. This is visible in topic 5, where the letters marked with square brackets are manually added afterwards for readability. Most of the topics are hard to interpret to coherent and understandable topics. Topic 7 has a some financial focus, which can be used to create a label. Similarly, topic 8 contains terms related to digitization and topic 5 focused on employment related skills and titles. Topic 6 contains innovation and data related terms and could be classified broadly under innovation topic. Other topics are hard to categorize. Many of the topics share key words making the topic definition hard. Hence, in conclusion, the modeling process has not been able to produce expected nor suitable for the defined use-case due to these reasons.

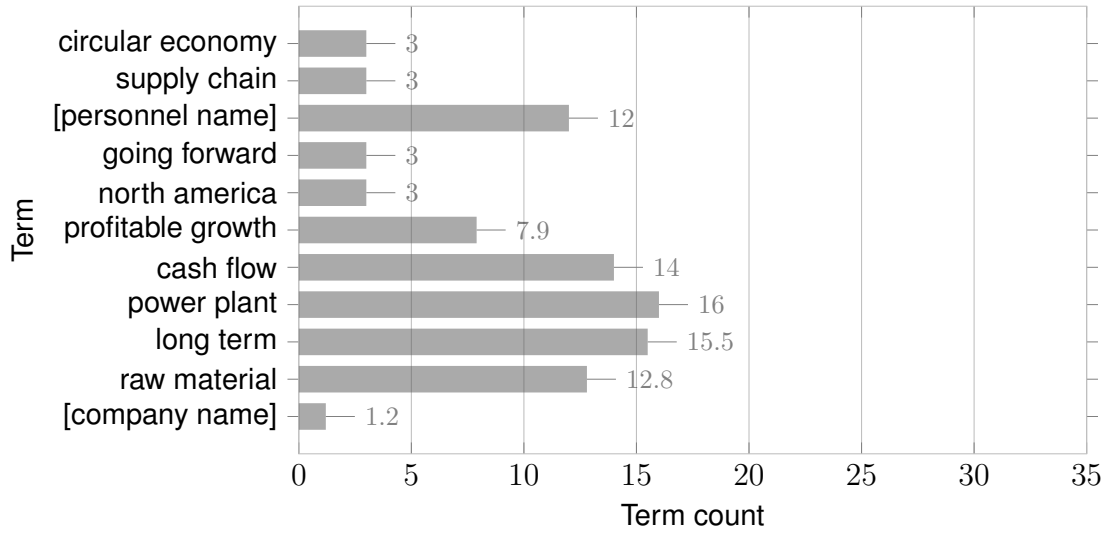


Figure 4.7. *Top 10 most common bigrams, sector dataset. Company and personnel names are redacted for privacy reasons*

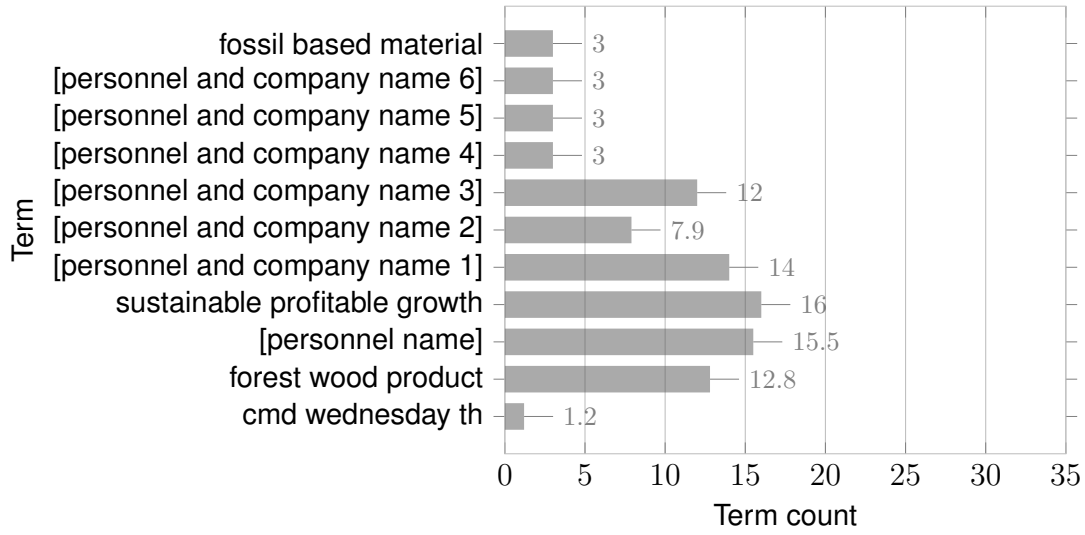


Figure 4.8. *Top 10 most common trigrams, sector dataset. Company and personnel names are redacted for privacy reasons*

Selected results for the sector dataset modeling are presented in Table 4.3. Few of the topics are cohesive and easily interpreted to a human, but majority seem arbitrary. To further analyse the contents of the topics the following classifications are defined for each topic. The topic number 9 contains words related to forestry and more specifically words related to forestry products. Topic 6 can be defined to focus on marine industry and shipping, while topic 13 is related to cargo handling and logistics in general. Good example of topics that are hard to indefinitely classify are topics number 2 and 5. Topic number 2 contains words related to recycling and circularity, but it also has terms that seemingly have no further connection to the overall theme. The same effect is visible in topic 5, which contains financial attributes, but similarly has terms that do not have a clear connection to financial matters. Finally, majority of the topics are hard to classify under a single theme.

Topics 1 and 15 are good examples of this and contains seemingly random terms that are hard to classify under a single theme.

Table 4.2. *Top 10 words of each topic, company dataset*

Topic name	Topic number	Top words
undefined	1	growth, modernization, asia, pacific, north, flow, , america, average, europe, expected
undefined	2	china, city, tier, growth, base, seeing, brand, number, growing, shanghai
undefined	3	growth, continue, quite, important, thing, china, , next, little, many, come
undefined	4	engineer, modernization, london, base, lift, safety, data, lead, task, contract
employment	5	manager, installation, responsible, skill, ethical, [bal]anced, seek, [ex]periences, [inno]vator
innovation	6	innovation, data, technology, important, partner, company, change, different, already, thing
finance	7	china, first, pricing, growth, europe, material, impact, asia, company, margin
digitalization	8	digital, capability, partner, startegy, connected, class, head, physical, create, different
undefined	9	growth, america, north, modernization, construction, thing, positive, productivity, pricing, growing

Table 4.3. *Top 10 words of selected topics, sector dataset*

Topic name	Topic number	Top words
undefined	1	electricity, ovat, kest, blog, rist, kanssa, programme, battery, hydro, joka
recycling	2	food, fiber, circular, flexible, ambition, india, carton, recyclable, circle, circulareconomy
finance	5	steel, cold, import, mill, ebitda, debt, ebit, cash, underlying, heavy
shipping	6	engine, vessel, offshore, installed, marine, load, ship, trade, book, flexible
forestry	9	forest, [company name], mill, course, pulp, wood, food, paperboard, chemical, fiber
logistics	13	terminal, automation, software, container, port, automated, cargo, margin, intelligent, ship
undefined	15	winter, factory, testing, russia, truck, road, startup, heavy, season, podcast

4.4 Dominant Topics and Topic Distributions

To better understand the underlying connections and contents of a topic, visualizations of top key words and corresponding word weight and count within a topic, and the topic distribution over the used corpora are inspected.

The document-topic distribution is not balanced in the modeling results. The figure 4.9 shows the document count per topic of the Company corpus, which contains documents from an individual company. The results show the high dominance of Topic 1. The Topic 1 includes 79.6 % of the documents and Topic 2 includes 12.0 % topics. Rest of the topics include 0–4 documents. This underlines the issues with the modeling process and the method's ability to provide deeper insights on the collected documents.

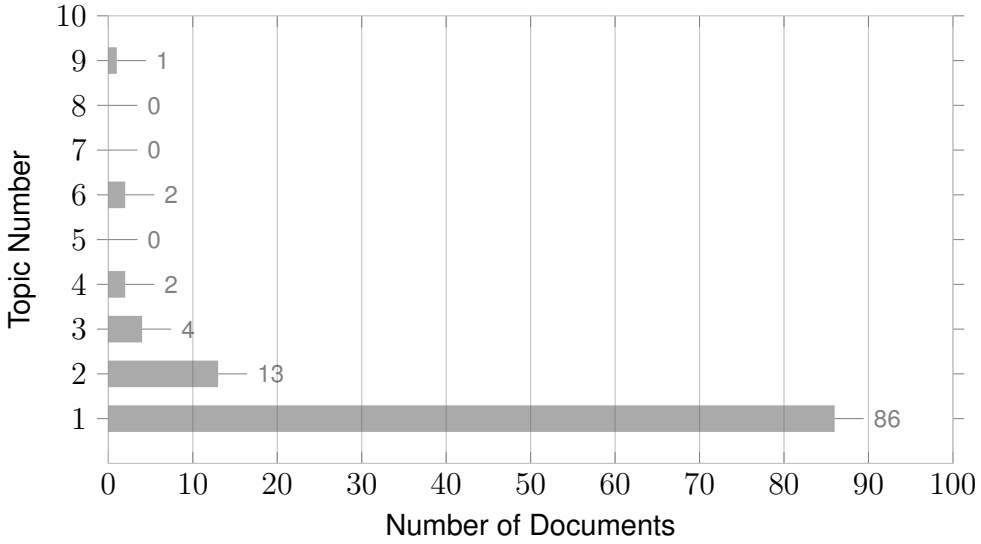


Figure 4.9. *Number of documents per topic, company dataset*

Topic 1 includes most of the documents in the corpus, but has an *undefined*-label in Table 4.2, due to the inability to determine a cohesive theme for the topic during human-assessment due to the randomness of the contents. The inability to label the topic halts the aim to produce additional information from the dataset. Similar issues are present with Topic 2, which holds the second-most documents. The topic is similarly labeled as *undefined* in Table 4.2 due to the lack of coherence during human assessment. This further supports the previous note on the challenges with the model's performance and ability to provide additional insights on the corpora.

One of the key interests in the work are the words that define a topic, since the aim is to be able to generate additional insights on the topics and themes discussed in the source material. The topic contents help to better understand the underlying connections within a topic and the terms that have the greatest impact on topic definition.

Figures 4.10, 4.11, and 4.12 present the dominant words and corresponding word-weights

for Topics 1, Topic 2, and Topic 5. In general, all the terms in all topics have relatively low counts and weights. The chosen Topics 1 and 2 include most of the Company-dataset documents and Topic 5 has the highest coherence based on human assessment. The exact counts of documents per topic are shown in Figure 4.9. The topic-word and word-weight graphs for all topics are presented in Appendix C.

The dominant words and word-weights for Topic 1 are visualized in Figure 4.10. The term '*growth*' has a high count and weight on Topic 1 compared to other terms in the Topic 1. Otherwise the terms have similar weights and counts providing a balanced topic. For Topic 2 the dominant words and word-weights are visualized in Figure 4.11. The Topic 2 has more variation with word counts, but the word weights are balanced. This creates a good topic, but the terms remain too random for analysis or additional insight purposes. It is also noted that the term 'growth' is present in both Topic 1 and Topic 2, which may cause confusion during document labeling. However, the term has a significantly higher weight in Topic 1 than in Topic 2.

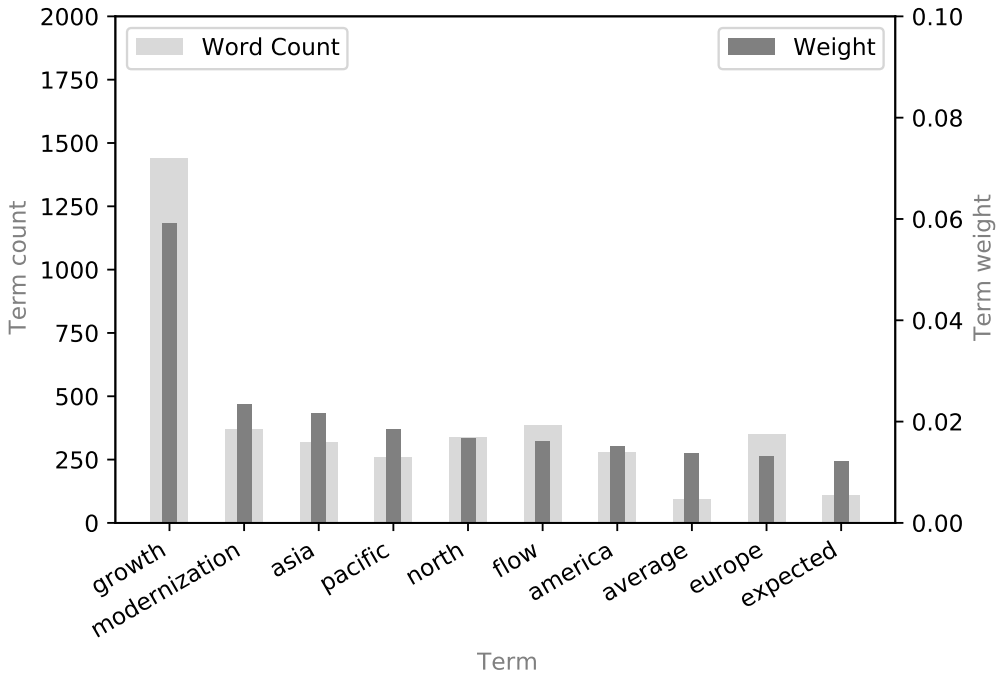


Figure 4.10. *Dominant words and word-weights, Topic 1, company dataset*

Topic 5 has been determined to contain *employment* - related terms and the dominant words and word weights are visible in Figure 4.12. The word counts for the top 10 terms and term weights are balanced, and the terms that define the topic are sensible to human assessment. However, both the word count and the word weight are relatively low. All of the word counts are under 250 and the weight is below 0.05 with each term. The topic is balanced, but the low weight for each term and the incomplete words (periences vs. experiences) makes the document classification difficult. As a result the modeling process labels zero documents under Topic 5.

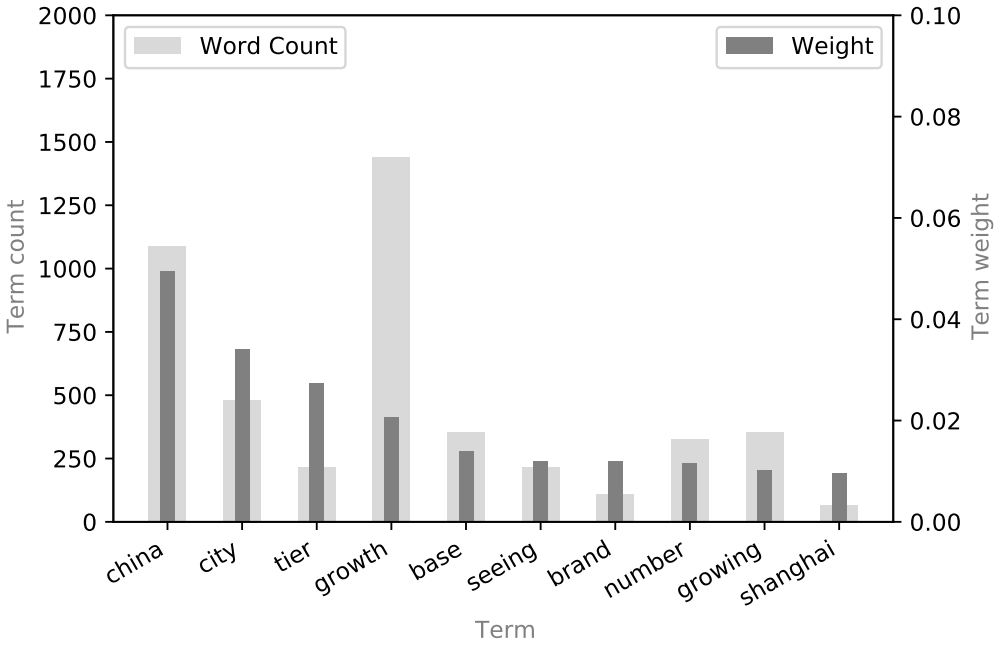


Figure 4.11. *Dominant words and word-weights, Topic 2, company dataset*

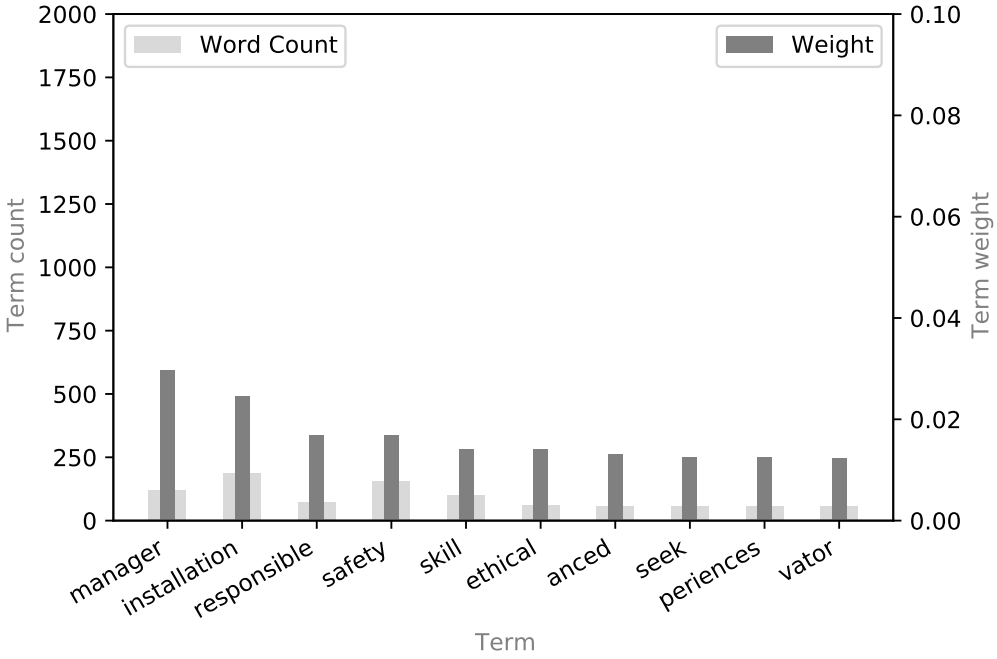


Figure 4.12. *Dominant words and word-weights, Topic 5, company dataset*

4.5 Qualitative Evaluation

The qualitative evaluation focuses both on the quality and usability of the system and the quality and usability of the analysis pipeline results. The experts assessing the results and the details of the qualitative evaluation methods and metrics are presented in Section 3.2.3.

4.5.1 System evaluation

The usability of the system is important during the development because the system is aimed to be used by people without extensive technical knowledge, i.e. the sales professionals. The preprocessing and analysis tools are used from the browser without any requirements to the local machine, which made the usage easy. The Google Drive based data storage is simple to use and includes a visual user interface. The authentication for to access data and run the code is done by a access prompt and using a login from the users Google account. However, the Vision API requires specific credentials for usage. The credential can be shared to other users in the same organization and thus the user does not need to obtain them themselves. In general, the system was seen as easy to use with limited technical knowledge. However, any error situation in authentication or running the code would require developer effort since the error messages in general are not intuitive.

4.5.2 Result evaluation

The focus is on n-gram search results presented in Figures 4.1 – 4.6 and the topic modeling results presented in Tables 4.2 and 4.3. The results are evaluated in three parts, with first part discussing the n-grams, the second part discussing the topic modeling, and the final part evaluating the overall results of the research based on the evaluation criteria-

The n-grams are computationally simple and efficient way to gather insights from a text set. The method does not contain any additional intelligence, nor require input from the user making the usage easy. The results describing the most occurring terms and occurrence counts are easy to understand and do not require any additional interpretation from the user. All of the n-grams produced contained search results that could be directly utilized during the sales process without any further analysis or modification. The results clearly present the most used terms, which allow the used to draw conclusions which topics may be important to the company in question. An expert with prior background knowledge is able to easily choose surprising or unexpected topics and terms from the n-gram results to be used as a conversation starter with a potential customer. Similarly, the results are easily presentable and interpretable to the customer. The n-gram findings can be utilized as a background and proof in fact-based arguments during the discussion. This brings credibility to the process, as the topics used for the approach can be supported with research findings and data analysis. However, the amount of company names, personnel information, and other cluttering terms in the results was high. This reduces the overall quality of the results, as the interest is on the topics discussed rather than on basic information of the company or the personnel. This issue was especially present with higher number of words n.

The topics created by the topic model algorithm resulted in only a few cohesive topics that can be easily interpreted or classified under a single theme by a human. The resulting topics require some further clarification and explanation prior to usage in sales environment. A minimal further processing includes labeling the topics manually with a human-defined names. The quality and understandability of the produced topics varied. The collections of words in a topic were often obscure and topics contained seemingly random terms without cohesive or interpretable meaning. However, the topics that had high coherence provided good insight into the topics present in the dataset and the keywords related to that theme. The trustworthiness of the results was halted by the low quality, which caused confusion. Nevertheless, overall the results were seen reliable enough to be used in a sales setting.

Overall, the coherence issues with both the n-gram search and the topic modeling approach greatly haltered the main effort to support the sales process with additional insights. The analysis tool was able to offer actionable insights only on few occasions, mainly due to the ambiguity of the results. However, both of the results were seen trustworthy enough to be used for backing arguments and statements during the sales process, when the results were clear enough to allow it.

5. DISCUSSION

The research conducted can be split into technical foundation and analysis. The aim of the first part the research, technical foundation, was to build a technical setup to automate and simplify text preprocessing for individuals lacking extensive technical knowledge, i.e. sales personnel. The pipeline successfully offers an easy way to clean, format, and aggregate data from different sources, as the pipeline automatically handles reading the data from storage, preprocessing, and saving the preprocessed text to storage. The technical setup built requires minimal human input and succeeds in automating the preprocessing task. The pipeline in general produces good results, but struggles to remove for example personnel names, company specific details, and common verbs, such as "said" and "see". The preprocessing pipeline can be further developed and adjusted relatively easily for future needs as each step is a separate block that can be easily modified, extended, or removed. The preprocessing sequence outputs a preprocessed corpora that is further used in the following research steps.

The second part of the research, data analysis and insight discovery, was the main focus of this work. The built a topic modeling approach discovers company related insight to work as a base for conversation starting during the customer approach phase in sales process. The chosen algorithm for the modeling was Latent Dirichlet Allocation, which is one of the most widely used topic modeling methods. The modeling process from the technical side is functional and suitable for the chosen use case. The process is easy to start and does not require extensive technical knowledge making it easily accessible to people without extensive technical knowledge. However, the chosen method of topic modeling was not able to create a modeling process that would produce coherent results. This is likely due to the low quantity and quality of the source data. The modeling produced, at best, a few easily interpretable topics, which were not enough to sustain the defined use-case.

The time spent running the modeling and analysing the results exceeds the following benefit in many cases, especially with time-limited sales people as the main user group. Similarly, the topic modeling results are often too ambiguous and require intensive further analysis. It become difficult to determine concrete talking points, which was the intended use-case, due to the ambiguity of the results. Referring back to the original research question defined in Section 1.1, the approach was able to only lightly and in few occasions

support the conversations initiated during the sales process. This is further supported with the notion of the n-gram search results being seen more useful as they were able to provide more concrete information of the most occurring topics in a format that is easier to interpret requiring less time to analyse.

In hindsight, the choice of method may not have been optimal for the selected use-case, as the goal was to get deeper insights of a dataset, but topic modeling in provides a general labeling of topics present rather than detailed information regarding the corpus. Further recommendations for future approaches are presented in Section 7.

5.1 Limitations

This study naturally has limitations and shortcomings. During the research process multiple challenges were faced, which were related mainly to the quality, quantity, and availability of the data. It is also important to note that the assessment method is heavily based on relatively few expert opinions, which may cause the final findings to be influenced by the experts opinions, prior experiences, or other similar matters rather than being an objective truth.

The datasets used for topic modeling were not optimal and the quality of data in the datasets varied a lot. The amount of poor, messy data affects the classification results negatively. There are multiple reasons for the poor quality of data. The documents used are primarily targeted to different stakeholders to share information regarding the company. The press releases are very marketing oriented and often do not reflect the actions of the company too closely. Similarly, the capital market day materials are aimed at current and potential investors and hence contain marketing related content. The lack of relevant information causes the analysis to become ineffective, as the information is polluted with noise created by marketing needs. As collected documents are merely for public use, they contain template information of the company, for example a company introduction at the end of a job posting. This information is not relevant for analysis purposes and is required to be removed. Finally, the documents provided by companies are sometimes extremely visual, causing the text to be difficult to read for a machine. A human would intuitively know in which order the text should be consumed, but a machine reads the text line by line horizontally regardless of the layout of the original text. This causes some context to be lost as the sentence structure or word order does not always remain the same. This is not an issue with topic modeling, but for the n-gram search the word order is crucial. Hence, the chosen approach may have missed elements of the text.

It is also assumed that the data gathered and used for the research contains the information this work aims to extract. The research is based on an assumption the companies talk about the most important matters in their materials, and hence the material can be used to determine the talking points and aspirations of a company. However, the is no

proof to back the argument. The materials are marketing material primarily targeted to the public and aim to create a favourable image of the company. Companies likely have future aspirations that are not shared in the materials due to for example sensitivity or controversy. For example, it may be business-wise a good decision to focus more on fossil fuels, but that would not be shared openly as it would likely cause controversy in today's climate concerned society.

Due to the generally low quality of the source text it required extensive preprocessing. The preprocessing sequence used successfully removes most of unwanted words and characters. Especially the list of stop-words to be removed is quite extensive due to the amount of clutter words, company, and personnel names in the documents. However, the amount of preprocessing also removes some words that could hold significant information value and hence hampers the final modeling result quality.

Some limitations are also noted on the result assessment. The assessment of results is heavily based on personal experience of experts. During to the relatively short research period, it was not possible to generate enough sales cases to do A/B testing or similar to verify the performance change. Hence, the study lacks quantitative metrics to determine the effect of the conducted work on the success of the sales process. The analysis to determine the usability of the final results is done by individual people, who are influenced by their opinions, surroundings, and prior knowledge. The research work conducted also determined that the most beneficial results are surprising or unexpected, as they work best as conversation starters. However, these definitions are highly dependant on personal experience and the present use-case. Hence, the assessment of the usability of the results is not objective, but subjective metric influenced by outside matters.

6. CONCLUSION

Consistent and successful sales flow is a key requirement in consulting business, where the profitability of the business depends on incoming client projects. The sales flow is a multi-step process beginning with identifying potential customers, qualifying the leads, drafting offers, and ending with a business agreement between parties. The focus of this work is on the customer approach phase. The sales flow begins with lead acquisition, where potential customers are identified. It continues with lead qualification, during which the potential customers are ranked based on internal criteria, such as company size or likelihood of becoming a customer. During the customer approach phase, the customer is commonly contacted for the first time regarding the new sales case. The approach phase focuses on starting a conversation with the potential customer rather than selling a product or a service instantly. Additional analysis of the current talking points from material published by the company allows the sales people to gain additional insights into the company and use the findings for the conversation-starting during the customer approach phase.

This thesis hypothesized that the materials published by companies could be used to provide insights into companies' business focus and aspirations. The goal of this study was to determine if and how topic analysis can augment the customer approach phase during the sales process. The aim was to discover conversation-starting topics by analyzing documents published by chosen companies. Based on the findings, the salespeople can offer targeted services in areas the potential customer is for example lacking compared to others in the same sector, or showing interest without current action to obtain the goal. Furthermore, the analysis gave the salespeople a data-based approach to support their arguments and claims during the process which increased credibility. All the material utilized is openly available online and utilized documents include the following: capital market day - materials, job postings, and press releases. The companies referred to in this study are picked from the clientele of the case company and therefore confidential.

The results obtained from this thesis research are in many cases insufficient to provide exact topics to be used as a conversation starter. The results obtained from the modeling method are too vague for the chosen use case, and consistently contain too much noise to be usable. This causes the work to positively impact the customer approach during the sales process only in a few cases. Furthermore, the topic modeling results always

require additional interpretation from the salespeople and can not be utilized directly after the analysis process. The additional analysis is both laborious and complex for people lacking technical knowledge opposing the usage of the built system. However, compared to the topic modeling results the n-gram search results are easy to interpret and often usable, but fail to gather deeper insights from the material. In conclusion, both methods provide a scientific and data-driven framework to obtain insights from the material. The analysis results can be used to support arguments and spark conversations during the customer approach only after an additional, manual assessment process, and provide sufficient insights only in few cases.

Both n-gram search and topic modeling suffer from the quality and quantity of available data. The used text data contains a lot of topic-specific stopwords that are hard to remove without extensive manual work, but often dominate the results, as they are present in the corpus with a high frequency. The manual data collection process is time-consuming and companies only produce a limited amount of documents within a reasonable analysis - time frame. The low absolute number of documents restraining the overall amount of documents available for analysis. The low amount of available text data negatively affects the quality of topics the Latent Dirichlet Allocation algorithm can produce.

This work concludes stating the designed analysis approach is successful in identifying conversation-starting talking points to augment the sales process only in a few cases. The most significant benefit the approach offers is a quantitative metric to back arguments regarding the most mentioned topics. The technical approach including the preprocessing pipeline, n-gram search, and topic modeling is successful and provides a good base for further development.

7. FUTURE WORK

In the final chapter of this thesis, proposals for future work applications suggested by the prior work are presented. The section propose ways to improve and further develop the previously presented research process, and discusses alternative approaches to sustain the research goal.

The work conducted in the scope of this thesis faced time constraints especially with manual data collection. A significant portion of the thesis time allocation was spent on manual work to acquire data. Speeding up the data collection process would alleviate a major bottleneck in the overall process. A sufficient amount of data is crucial for the success of the analysis process. Automating the data collection process for future work is suggested as it would significantly speed up the process and allow the experts to better focus on the relevant parts of the work rather than to time-consuming manual data acquisition.

The low amount of source text was one of the biggest challenges during the research work. To alleviate this, alternative data sources should be applied. For the sector dataset, industry reports can be used as a data source to have a wider understanding of the sector focus and aspirations. A wide selection of reports is freely available online from different industries and geographical locations. For example, the International Energy Agency [1], United Nations Economic Commission for Europe [2], and large consulting companies publish reports discussing the current status and future developments of the industry. One possibility for better quality datasets could be obtaining ready-made sector-related datasets, but those commonly require payment and hence are not optimal for the continuous analysis process.

Investigation of alternative preprocessing methods for the source texts to improve the current preprocessing pipeline are also proposed. Future research would benefit from the utilization of a more effective preprocessing schema, which leads to results with better quality and higher coherence. The texts used contain a significant amount of clutter, such as filler words, names, and boilerplate descriptions that are regularly present but have low information value.

The selected topic modeling approach treats all words with the same significance caus-

[1] https://www.iea.org/data-and-statistics
[2] https://unece.org/publications/oes/welcome

ing the topics to be noisy and dominated by high-frequency, but often meaningless words. Some earlier research suggests the quality of topics produced by the model could be improved by additional processing, such as assigning weights to terms before modeling [32]–[34]. Entropy-based term weighting (EW) introduced in [32] assigns weights to individual words based on the influence of the occurrence of other words. A higher influence on the occurrence of other words is denoted as higher weight. The method works well for both short and long texts, which is ideal for this use case, as the texts modeled have varying lengths. Wilson and Chew (2010) in [33] describe two weighting methods using a logarithmic function that assigns smaller weights to high-frequency words that are assumed to convey little useful semantic information. Term weighting LDA (TWLDA) defines domain-specific stop-words scattered across most topics using a BDC weighting schema and punishes such terms decreasing their effect on final results [34]. All of the methods are shown to outperform traditional LDA with uniform weighting in topic coherence and improve the generated topics.

REFERENCES

[1] J. Lyly-Yrjänäinen, T. Mahlamäki, T. Rintamäki, H. Saarijärvi, and V. Tiitola, *Sales in Technology-driven Industries*. Technology Industries of Finland, 2019, ISBN: 978-952-238-230-6.

[2] P. Espadinha-Cruz, A. Fernandes, and A. Grilo, "Lead management optimization using data mining: A case in the telecommunications sector", *Computers Industrial Engineering*, vol. 154, pp. 107–122, 2021.

[3] J. D'Haen and D. Van den Poel, "Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework", *Industrial Marketing Management*, vol. 42, no. 4, pp. 544–551, 2013.

[4] J. D'Haen, D. Van den Poel, D. Thorleuchter, and D. Benoit, "Integrating expert knowledge and multilingual web crawling data in a lead qualification system", *Decision Support Systems*, vol. 82, pp. 69–78, 2016.

[5] M. Cooper and C. Budd, "Tying the pieces together: A normative framework for integrating sales and project operations", *Industrial Marketing Management*, vol. 36, no. 2, pp. 173–182, 2007.

[6] M. Skaates and H. Tikkanen, "International project marketing: An introduction to the inpm approach", *International Journal of Project Management*, vol. 21, pp. 503–510, 7 2003.

[7] R. D. Wilson, "Using online databases for developing prioritized sales leads", *Journal of Business Industrial Marketing*, vol. 18, pp. 388–402, 2003.

[8] N. Syam and A. Sharma, "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice", *Industrial Marketing Management*, vol. 69, pp. 135–146, 2018.

[9] J. Paschen, M. Wilson, and J. J. Ferreira, "Collaborative intelligence: How human and artificial intelligence create value along the b2b sales funnel", *Business Horizons*, no. 3, pp. 403–414, 2020.

[10] H. Jelodar, Y. Wang, C. Yuan, *et al.*, "Exploring prevalence of wound infections and related patient characteristics in homecare using natural language processing", *Multimedia Tools and Applications*, vol. 78, pp. 15 169–15 211, 2021.

[11] B. Chae and E. Park, "Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling", *Sustainability*, vol. 10, no. 7, p. 2231, 2018.

[12] K. Woo, Song, L. Currie, *et al.*, “Exploring prevalence of wound infections and related patient characteristics in homecare using natural language processing”, *International Wound Journal*, vol. 19, no. 1, pp. 211–221, 2021.

[13] N. Pröllochs and S. Feuerriegel, “Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling”, *Information & Management*, vol. 57, no. 1, p. 103 070, 2020.

[14] M. Blei, “Probabilistic topic models”, *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[15] J. Lau, D. Newman, and T. Baldwin, “Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality”, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, 2014.

[16] M. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[17] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing semantic coherence in topic models”, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272, 2011.

[18] D. Jurafsky and J. Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd edition.* 2021.

[19] M. Nelimarkka, “Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: Kriittisiä havaintoja”, *Politiikka: Valtiotieteellisen yhdistyksen julkaisu*, vol. 1, no. 61, pp. 6–33, 2019.

[20] D. Greene, D. O’Callaghan, and P. Cunningham, “How many topics? stability analysis for topic models”, in *Machine Learning and Knowledge Discovery in Databases*, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds., Springer Berlin Heidelberg, 2014, pp. 498–513.

[21] D. Maier, A. Waldherr, P. Miltner, *et al.*, “Applying lda topic modeling in communication research: Toward a valid and reliable methodology”, *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 93–118, 2018.

[22] F. Lind, J.-M. Eberl, O. Eisele, T. Heidenreich, S. Galyga, and H. G. Boomgaarden, “Building the bridge: Topic modeling for comparative research”, *Communication Methods and Measures*, vol. 16, no. 2, pp. 96–114, 2022.

[23] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures”, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408, 2015.

[24] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence”, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 100–108, 2010.

[25] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models”, *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112, 2009.

[26] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, “Reading tea leaves: How humans interpret topic models”, *Advances in neural information processing systems*, vol. 22, 2009.

[27] F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both, “Evaluating topic coherence measures”, *arXiv preprint arXiv:1403.6397.*, 2014.

[28] D. Newman, S. Karimi, and L. Cavedon, “External evaluation of topic models”, *Proceedings of the 14th Australasian Document Computing Symposium*, pp. 11–18, 2009.

[29] G. Bouma, “Normalized (pointwise) mutual information in collocation extraction”, *Proceedings of GSCL*, vol. 30, pp. 31–40, 2009.

[30] A. Abuzayed and H. Al-Khalifa, “Bert for arabic topic modeling: An experimental study on bertopic technique”, *Procedia Computer Science*, vol. 189, pp. 191–194, 2021.

[31] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, “Exploring topic coherence over many models and many topics”, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961, 2012.

[32] L. Ximing, Z. Ang, L. Changchun, O. Jihong, and C. Yi, “Exploring coherent topics by topic modeling with term weighting”, *Information Processing Management*, vol. 54, no. 6, pp. 1345–1358, 2018.

[33] A. Wilson and P. A. Chew, “Term weighting schemes for latent dirichlet allocation”, *In: human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 465–473, 2010.

[34] K. Yang, Y. Cai, Z. Chen, H.-f. Leung, and R. Lau, “Exploring topic discriminating power of words in latent dirichlet allocation”, *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2238–2247, 2016.

APPENDIX A: MANUALLY CURATED LIST OF STOP-WORDS

A manually curated set of stops words is used additionally during preprocessing to remove unnecessary noise and better the quality of the source text for analysis. Some company and personnel names are missing from this list due to privacy concerns, but it is recommended to add at least the names of the case companies.

```
extended_stopwords = ['q', 'a', 'b', 'c', 'f', 'j', 'o', 'r', 'd',
'e', 'n', 'r', 'l', 'fj', 'ds', 'u', 'h', 'p', 'v', '2',
'1', 'sw', 'us', 'yo', 'rs', 'ac',
'am', 'pm' 'ect', 'new', '-', 'jt', '20', '2020', '-', '0', '1',
'2', '3', '4', '5', '6', '7', '8', '9', '10', '-1', '-2', '-3', '-4',
'-5', '-6', '-7', '-8', '-9', '-10', '-20', '-30', '-40',
'manuel', 'roj', 'spokesperson', 'com', 'de', 'ly', 'http', 'bit',
'https', 'fal', 'cn', 'year', 'financial', 'report', 'million',
'group', 'statement', 'capital', 'market', 'day', 'eest', 'eur', ¨
'per','oyj', 'oy', 'gmt', 'co', 'llc','plc', 'ha','also', 'www',
' ee', 'en', 'eet', 'january', 'february', 'march',  'april', 'may',
'june', 'july', 'august', 'september', 'october', 'november',
'december', 'wa', 'meur','bulletin', 'unknown', 'corporation',
'speaker', 'look', 'like', 'posted', 'myworkdayjobs',  'president',
'vice','page', 'kind', 'english', 'job', 'search', 'makati', 'cer',
'gk','today', 'apply', 'omaha', 'ne', 'posted', 'ago',  'start',
'want', 'really', 'still', 'something', 'question', 'okay',
'myworkdayjobs', 'trademark' ]
```

APPENDIX B: SECTOR DATASET TOPIC MODELING RESULTS

Table B.1. *Sector dataset topic modeling results. Personnel and company names are removed for privacy reasons*

Topic name	Topic number	Top words
undefined	1	electricity, ovat, kest, blog, [personnel name], kanssa, programme, battery, hydro, joka
recycling	2	food, fiber, circular, flexible, ambition, india, carton, recyclable, circle, circulareconomy
paper industry	3	forest, wood, pulp, [company name], mill, course, paperboard, division, fibre, profitable
undefined	4	mill, pulp, automation, tissue, roll, webinar, internet, fiber, wastewater, measurement
finance	5	steel, cold, import, mill, ebitda, debt, ebit, cash, underlying, heavy
shipping	6	engine, vessel, offshore, installed, marine, load, ship, trade, book, flexible
undefined	7	pulp, course, mill, marine, engine, vessel, solar, wood, food, renewables
undefined	8	course, wind, solar, electricity, russia, utility, priority, flexibility, hydro, dividend
forestry	9	forest, [company name], mill, course, pulp, wood, food, paperboard, chemical, fiber
undefined	10	quite, course, margin, side, little, said, mining, couple, type, obviously
marine	11	engine, vessel, marine, storage, shipping, ships, solar, hybrid, renewables, grid
undefined	12	chemical, treatment, pulp, polymer, revenue, mining, ebitda, said, margin, capex
logistics	13	terminal, automation, software, container, port, automated, cargo, margin, intelligent, ship
undefined	14	mill, pulp, wood, forest, course, [company name], ambition, steel, ebit, [personnel name]
undefined	15	winter, factory, testing, russia, truck, road, startup, heavy, season, podcast

APPENDIX C: WORD COUNT AND WEIGHT OF TOPIC KEYWORDS PER TOPIC, COMPANY DATASET

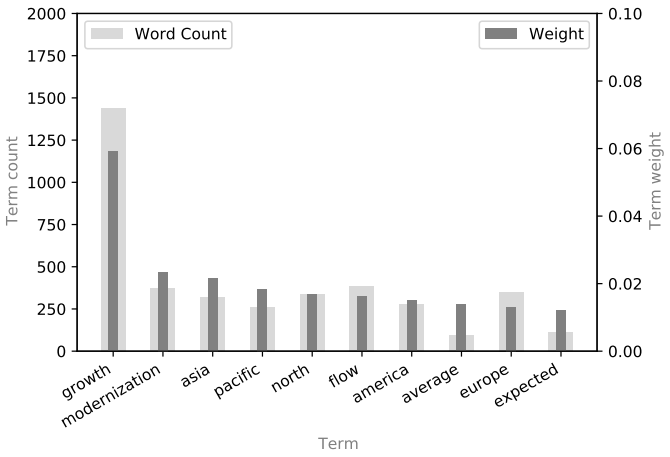


Figure C.1. *Topic 1, company dataset*

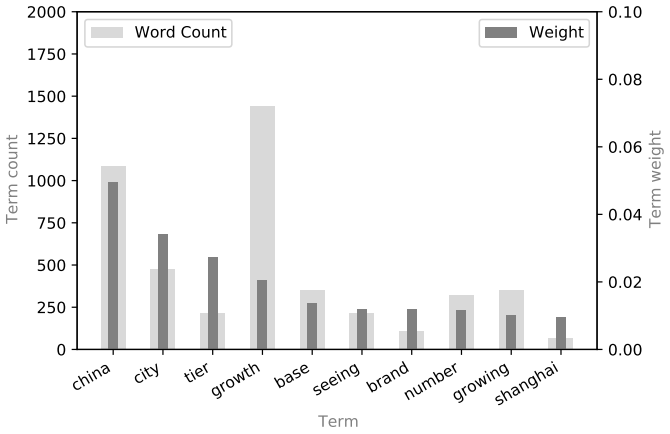


Figure C.2. *Topic 2, company dataset*

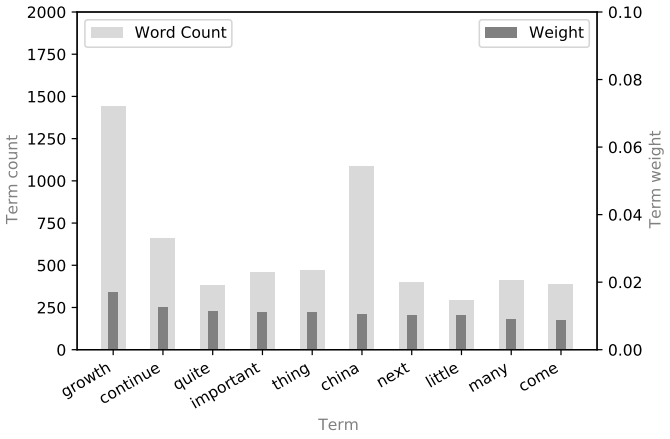


Figure C.3. *Topic 3, company dataset*

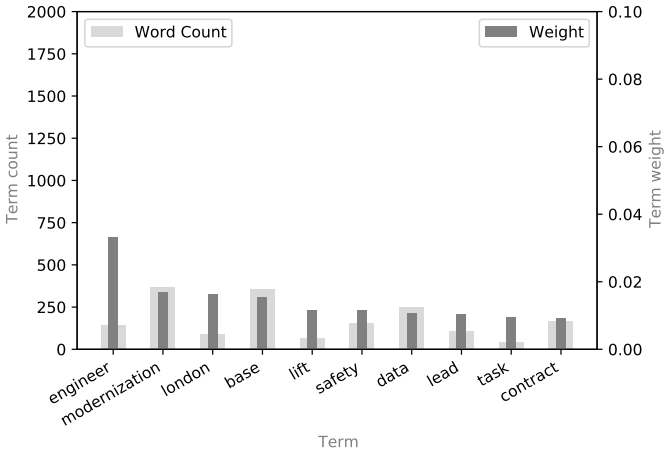


Figure C.4. *Topic 4, company dataset*

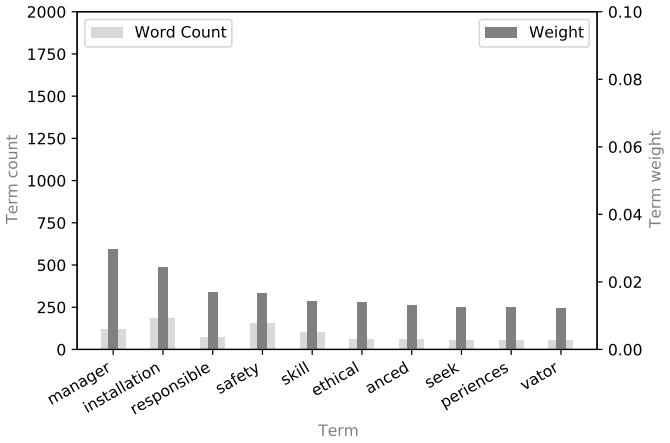


Figure C.5. *Topic 5, company dataset*

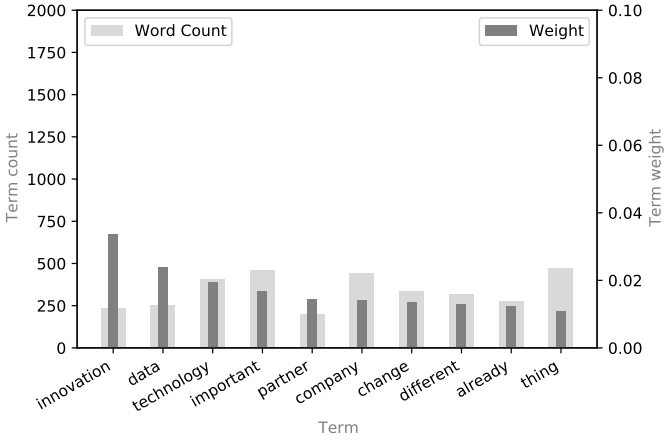


Figure C.6. *Topic 6, company dataset*

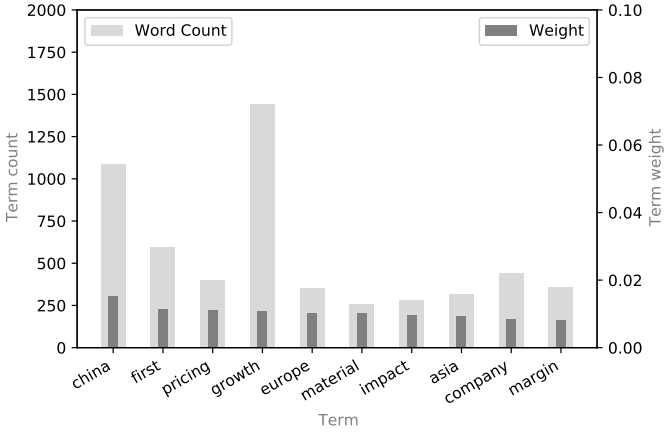


Figure C.7. *Topic 7, company dataset*

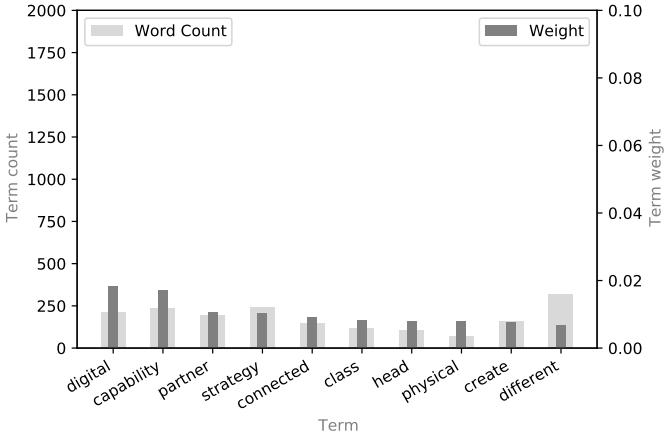


Figure C.8. *Topic 8, company dataset*

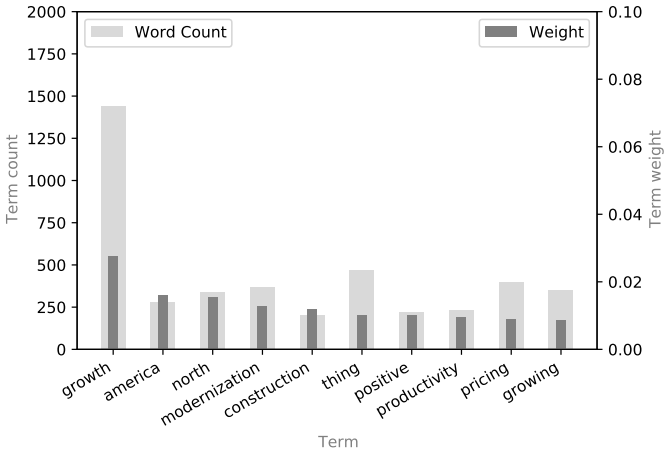


Figure C.9. *Topic 9, company dataset*