Master Thesis

Evaluating and applying machine learning algorithms to analyze entrepreneurial ecosystems

Chair: Entrepreneurship, Innovation and Technological Transformation

First supervisor: Prof. Dr. Dries Faems

Second supervisor: Maxim Mommerency

Vallendar, January 19, 2021

Derstappen, Raphael

20003020

14.06.1995 in Neuwied, Germany

Eichenweg 3

56427 Siershahn

Abstract

Entrepreneurial ecosystems are an engine of growth for national economies because interdependent actors enable productive entrepreneurship within this territory. Therefore, research on entrepreneurial ecosystems accelerated over the last years with a primary focus on their identification and measurement. However, research in this field largely neglected in-depth analysis of specific industries to assess the status quo of entrepreneurial ecosystems and innovation distribution. Hence, I try to close this gap by introducing a data-driven methodology that utilizes Natural Language Processing to analyze an entrepreneurial ecosystem.

Thereby, I borrowed algorithms from the Natural Language Processing subfield of topic modeling to introduce the techniques in a management research context. More specifically, I applied and evaluated three distinct algorithms, namely Latent Dirichlet Allocation, Contextualized Topic Modeling, and Correlation Explanation, to assess their usefulness in creating coherent and meaningful topics for further ecosystem analysis. To create the topics with unsupervised and semi-supervised machine learning, I leveraged a dataset on cybersecurity companies from Crunchbase and Startup Nation Central. I chose the cybersecurity industry as a showcase due to its increasing importance in an interconnected, digital world and its inherent fragmentation in several sub-categories. Moreover, I introduced a new approach by combining Correlation Explanation with Word2Vec to automate and more accurately find valuable anchor words that support Correlation Explanation in finding coherent topics.

The outcomes of the topic modeling algorithms illustrate the viability of the approach in analyzing entrepreneurial ecosystems. Although substantial differences in coherence and insightfulness of the topics based on the unsupervised and semi-supervised methods exist, they provide the basis for an initial understanding (through Latent Dirichlet Allocation and Contextualized Topic Modeling) as well as a more detailed understanding of an entrepreneurial ecosystem (through Correlation Explanation with Word2Vec).

Thus, I recommend researchers and practitioners to include this machine learning approach to extend their current entrepreneurial ecosystem research methods. While research in Natural Language Processing happens largely in the computer and data science domain, it should not keep researchers from the application in a management science context. With the constant innovation in Natural Language Processing, the prospects for a more in-depth and data-driven analysis of entrepreneurial ecosystems looks promising.

Keywords: entrepreneurial ecosystem, topic modeling, NLP, LDA, CTM, CorEx, startups cybersecurity

Table of contents

List of figures

List of tables

Abbreviations

NLP: Natural Language Processing

LDA: Latent Dirichlet Allocation

CTM: Contextualized Topic Modeling

CorEx: Correlation Explanation

VC: Venture Capital

IoT: Internet of Things

GDPR: General Data Protection Regulation

TF-IDF: Term-frequency inverse-document-frequency

BERT: Bidirectional Encoder Representations from Transformers

BMBF: Bundesministerium für Bildung und Forschung

SWOT: Strengths, Weaknesses, Opportunities, Threats

1. Introduction

Startups are catalysts for economic growth. With new ideas, services, products, and business model innovation, startups modernize existing economic structures and create new jobs (Audretsch, Keilbach, & Lehmann, 2006). In Germany alone startups create around 430,000 full-time jobs every year, and in the United States it is more than three million jobs (Bureau of Labor Statistics, October 2020; Federal Ministry for Economics Affairs and Energy, n.d.). Due to the importance of startups in the national economy, scholars started to research the factors driving entrepreneurship, which resulted in the emerging field of Entrepreneurial Ecosystems (Isenberg, 2011; Roundy, Bradshaw, & Brockman, 2018; Stangler & Bell-Masterson, 2015). While the main focus within this field is on identifying and measuring Entrepreneurial Ecosystems, an analysis of a specific industry on a national or regional level does not exist. Hence, this Master Thesis attempts to close this gap with a more data-driven approach to analyze Entrepreneurial Ecosystems by leveraging state-of-the-art machine learning algorithms.

A completely new approach to analyzing Entrepreneurial Ecosystems represents machine learning algorithms for topic modeling. Advances in natural language processing (NLP), which deals with processing and analyzing large amounts of natural language data with computers, make this new approach possible (Hannigan et al., 2019). Based on the text description of startups, I identified hidden topics with the algorithms Latent Dirichlet Allocation (LDA), Contextual Topic Modeling (CTM), and Correlation Explanation (CorEx) and assigned them to the respective startup. The hidden topics are nothing but a specific keyword describing the affiliation of the startup to a specific field within an industry. In this Master Thesis, the Cybersecurity industry is the focus.

I chose the Cybersecurity industry because of its increasing importance in the digital world, its fragmentation into several sub-categories and its switch from an IT problem to a business problem. In todays' world, cybersecurity plays a vital role for businesses, consumers, and governments due to the increased connectedness to the internet in every aspect of life (de Bruijn & Janssen, 2017). For example, a study by McKinsey (2019) projects the number of devices to be connected to the internet to increase to 43 billion by 2023 compared to just 14 billion in 2018. Nowadays, the internet of things is just an example of the connectedness to the internet and a single point of vulnerability for entities. In general, a large number of attack surfaces exist which makes companies often dependent on third-party solutions to enhance their cybersecurity (Accenture, 2020). These third-party solutions frequently spring from startups that quickly adjust to the fast-moving cybersecurity space.

Entrepreneurial Ecosystems are a national and regional phenomenon. Probably the most well-known entrepreneurial ecosystem is Silicon Valley in California. Additionally, Israel is well known for its high-tech startups and its affinity for cybersecurity startups in particular. Germany is a unique ecosystem because of its fragmentation throughout the country in Berlin, Munich, Hamburg, or the Rhein-Ruhr metropolitan region. Hence, I analyzed the Entrepreneurial Ecosystem in Europe focusing on Germany, Israel, and the USA to identify differences and similarities in the various verticals of cybersecurity.
For this purpose, I try to answer an evaluation and a comparative research question to add value to existing research on entrepreneurial ecosystems and to extend the current methodologies by focusing on natural language processing algorithms. The two research questions are:

1. *What are the advantages and disadvantages of topic modeling machine learning algorithms such as LDA, CTM, and CorEx for the analysis of entrepreneurial ecosystems? (Evaluation research)*
2. *Resulting from the best topic modeling approach evaluated in Research Question 1, what are the differences and similarities between Europe (particularly Germany), Israel, and the USA regarding their cybersecurity ecosystem?*

In order to answer these research questions, I divided this Master Thesis into multiple sections. First, a literature review sheds light on existing research on entrepreneurial ecosystems. More specifically, I describe the identification and measurement as well as the participating stakeholders in an entrepreneurial ecosystem. Second, I outline the methodology of this Master Thesis. Beginning with the provision of a context of cybersecurity and the different nations in focus, the methodology section clearly explains the process from data collection, data preprocessing and formatting, up to the machine learning algorithms in use as well as the approach of topic assignment. Third, I display the results. In particular, I compare the topic modeling outcomes from the different algorithms based on objective and subjective measures. Next, I examine the final topics as well as their distribution over the geographic locations. In the last part of the results section, I report the topic and startup distribution on a regional level to identify entrepreneurial hotspots within Europe, particularly Germany, Israel, and the USA. Fourth, I answer the research questions in greater detail based on the results section's outcomes, further research, and interviews with industry experts. I accompany this discussion section with the limitations of the methodology and a direction of future research. Last, I conclude this Master Thesis ends with a summary of the main outcomes of this research.

2. Literature review

The literature review focuses on the previous findings of scholars concerning the entrepreneurial ecosystem. Thereby, I provide a comprehensive overview of existing "classic" and state-of-the-art literature to find a suitable definition of the entrepreneurial ecosystem. Subsequently, I dive deeper into the definition to distinguish productive entrepreneurship and the changing perspective of entrepreneurs in today's academic world. Afterward, I present an overview of the various stakeholders and their importance to the entrepreneurial ecosystem. Last, I display different frameworks and theories to identify and measure entrepreneurial ecosystems before evaluating the current state of research.

2.1 The Entrepreneurial Ecosystem

The discussion of what encompasses an entrepreneurial ecosystem and how to measure it is still ongoing in the academic world. To grasp the overall concept of an entrepreneurial ecosystem, I looked at the two terms in isolation. First, being entrepreneurial means "exploring, evaluating, and exploiting opportunities for creating new goods and services" (Cavallo, Ghezzi, & Balocco, 2019, p. 5). Second, the ecosystem approach comes from the biological science field in which an ecosystem comprises a community of interacting organisms and their physical environment. Academic papers frequently point out that ecosystems within the social science and management context spring from an analogy of the biological ecosystem (Cavallo et al., 2019; F. Stam & Spigel, 2016). After Moore (1993) introduced the business ecosystem in the management literature, other researchers created offspring's, such as university-based ecosystems (Rice, Fetters, & Greene, 2014), organizational ecosystem (Mars, Bronstein, & Lusch, 2012) and innovation or knowledge ecosystems (Clarysse, Wright, Bruneel, & Mahajan, 2014; Jackson, 2011). In recent years, the focus of scholars has shifted further to the entrepreneurial ecosystem with the entrepreneur and interdependent other actors at its heart. From this standpoint, scholars established a more fundamental definition and understanding of the entrepreneurial ecosystem. In the following, solely the entrepreneurial ecosystem is in focus with its key element of productive entrepreneurship.

2.1.1 Definition of the entrepreneurial ecosystem

Analyzing the role of entrepreneurs and startups concerning innovation and economic welfare is something scholars agree on. However, a global definition of the term "entrepreneurial ecosystem" does not exist so far. Cavallo et al. (2019) underline the discussion of scholars by identifying more than 15 different definitions of an entrepreneurial ecosystem in their paper.

Although scholars have not accepted a single definition, all of the definitions seem to link to dimensions such as *interdependence, interconnectedness, economic growth, productive entrepreneurship,* and *regional context.*

The entrepreneurial ecosystem approach focuses on the external environment similar to established concepts such as clusters, industrial districts, or innovation systems. Clusters are "geographic concentrations of interconnected companies, specialized suppliers, service providers, firms in related industries, and associated institutions [...] in particular fields that compete but also co-operate" (Porter, 1998, p. 197). Within the industrial district, the focus is on co-operation and competition between firms that somewhat establish their local economy by producing and buying within their respective geolocation (Becattini, 2004). The innovation system approach deals with knowledge spillovers among different organizations within a region which increases overall innovativeness, often initiated between universities or research institutes and innovative firms (Cooke, Uranga, & Etxebarria, 1997). All three concepts deal with the connection of firms within a regional area and the larger system of innovation by focusing on larger companies with more established processes and related industries while neglecting the role of entrepreneurs. In comparison, the entrepreneurial ecosystem treats entrepreneurs and startups as unique entities with different resources and capabilities within the context of geolocational innovation (Nicotra, Romano, Del Giudice, & Schillaci, 2018).

Hence, for this Master Thesis, the definition of an entrepreneurial ecosystem will be "a set of interdependent actors and factors coordinated in such a way that they enable productive entrepreneurship within a particular territory" (F. Stam & Spigel, 2016, p. 1). This definition fully combines the themes of multiple interdependent actors, entrepreneurship focused on growth and innovation in a confined location.

2.1.2 Productive entrepreneurship

From this standpoint, it is important to understand what productive entrepreneurship encompasses. Productive entrepreneurship "..., refers, simply, to any activity that contributes directly or indirectly to net output of the economy or to the capacity to produce additional output" (Baumol, 1994, p. 30). More specifically, with the entrepreneurial ecosystem framework in mind, productive entrepreneurs create value for themselves but also for the economy at large. This differs from a previous belief that even self-employed are part of the productive entrepreneurship scheme, who might create value for themselves, but not for society at large. On the other spectrum, employees might create value for the organization and society while not being independent. Henrekson and Sanandaji (2014) identify that productive

entrepreneurship "reduces the small business share of employment" (p. 1764). They describe that entrepreneurs bring innovations to market which often lead to companies with thousands of high-paying jobs. People occupy these jobs who would otherwise work for themselves. This supports the notion that self-employed should be treated separately from the entrepreneurial ecosystem.

Productive entrepreneurship is important because it creates several economic outputs. According to Nicotra et al. (2018), productive entrepreneurship 1) creates jobs and reduces unemployment, 2) generates innovation and explores new markets, 3) commercializes new ideas or technologies, and 4) increases competition which boosts market efficiency and ultimately people's welfare. In literature, these types of firms, that are responsible for job creation and productivity growth, are called "high-growth firms" (Daunfeldt, Elert, & Johansson, 2014) and equated with the term startup.

2.2 Stakeholder analysis

The literature review revealed that the entrepreneurial ecosystem is particularly distinguished by its interdependent actors and factors that enable productive entrepreneurship. Thereby, entrepreneurs, investors, universities, corporates and political decision-makers are associated with the role of interdependent actors within the system. They all play an essential role by enabling entrepreneurship through idea execution, financing options, know-how, and providing the regulatory framework to operate. In the following, I describe the stakeholder's and their importance for the ecosystem.

2.2.1 Entrepreneurs

For the entrepreneurial ecosystem to properly work, the entrepreneur is the most critical part. Hence, it is necessary to understand what an entrepreneur is and what an entrepreneur is not necessarily. Schumpeter (1934) was one of the first researcher who worked on the influence of entrepreneurship on the economy and the role of the entrepreneur. He described entrepreneurs as individuals that use existent resources to "create new combinations and new uses". Thereby, the entrepreneur is seen as an actor that is highly relevant for economic development. Further research went on to distinguish entrepreneurs from small business owner and managers. A small business owner is defined as "an individual who establishes and manages a business for the principal purpose of furthering personal goals" (Carland, Hoy, Boulton, & Carland, 1984, p. 358). This is distinct from the entrepreneur whose main purpose for the business is profit and growth by utilizing strategic management (Carland et al., 1984). The characteristics of the

entrepreneur are also traceable in the previous described definition of productive entrepreneurship. In accordance with existing literature, Stewart, Watson, Carland, and Carland (1999) conclude, from an extensive survey study, that entrepreneurs have a higher motivation for achievement, a more risk-taking personality and a preference for innovation compared to corporate managers and small business owners.

This macro perspective of the identifiers of an entrepreneur have been further researched and frameworks were created to identify an entrepreneur based on character themes. Bolton, Thompson, and Thompson (2003) created the FACETS framework to define the entrepreneur based on six-character themes; focus, advantage, creativity, ego (inner and outer), team and social. Focus, advantage and creativity work in tandem, because creativity is needed to identify opportunities, whereas advantage leads to pursuing the most promising ideas and focus ensures an effective implementation. The inner ego deals with dedication and motivation to achieve something, whereas the outer ego creates the desire to be in charge and the willingness to deal with setbacks. The team component is a multiplier so that complementary people work together, entrepreneurs know when to ask for support and to understand the value of networking. The social dimension simply influences the direction of the business, whether it is for-profit or non-profit and the culture of the business. These themes or characteristics are recurrent when analyzing successful entrepreneurs. Nevertheless, the discussion whether entrepreneurs are born or made is not resolved. Hence, it is difficult to apply one framework to appropriately identify an entrepreneur. It is often not clear whether specific character traits of an entrepreneur have developed during the entrepreneurial journey or existed beforehand (Kerr, Kerr, & Xu, 2017).

Although, there is not an agreed-on mechanism to identify the entrepreneur, it is obvious that the entrepreneur plays the central figure within the entrepreneurial ecosystem. Without the entrepreneur, research & development would not be acted upon, leading to a plateau in economic development and consequently, a slowdown in the improvement of life and standard of living.

2.2.2 Investors

Entrepreneurial ventures, especially high-tech ones, are financially constrained. Without a track record of past success, credibility and reputation, as well as non-existence of tangible assets for collateral, it is difficult to gain access to debt capital from banks. Moreover, the existence of information asymmetry and high uncertainty, present in entrepreneurship, increase the barriers to access financing from more traditional investors (e.g. hedge funds, institutions) (Murphy &

Edwards, 2003). As a consequence, venture capital (VC) firms have emerged to bridge that gap. With a higher risk-profile than other investors, VCs participate in equity rounds to gain an equity share in the startup to realize a potentially substantial return in a future exit scenario (Grilli, Mrkajic, & Latifi, 2018). Therefore, VCs are especially important in the early stages of a startup and in follow-up rounds to inject new capital when necessary. After the startup becomes more established, it is necessary to have other sources of financing within the ecosystem. This is due to the structure of VCs that make it difficult to infinitely support their portfolio companies.

Often, the investors have been entrepreneurs themselves. Therefore, VCs offer more than monetary investment to startups. With their experience they can provide guidance, network and operational experience in the entrepreneurial ecosystem. However, evidence-based articles on VCs are rare and hence, it is still unsure how much non-monetary resources provide value for the economic growth for the startups (Amornsiripanitch, Gompers, & Xuan, 2019).

Overall, the role of VCs has become more important over the last years measured by the number of VCs incorporated and amount of capital injected. With less venture capital investments after the financial crisis in 2008/2009, investments have exceed previous levels by far since 2014 (OECD, 2020). Although, scholars point out that most VCs do not generate abnormal positive returns for their limited partners, it is undisputable that they play an important role in the entrepreneurial ecosystem (Mulcahy, Weeks, & Bradley, 2012).

2.2.3 Universities

According to Audretsch (2014), the role of university has changed when economic performance through physical capital and unskilled labor was replaced by knowledge as the driving force underlying economic growth. Since then, entrepreneurial universities emerged, which devote research to providing solutions for societal problems and challenges. Hence, universities act as important actor for knowledge spillovers within the entrepreneurial ecosystem which leads to the commercialization of ideas (Kantor & Whalley, 2014). Moreover, universities educate young people in numerous fields leading to high-skilled employees that contribute to economic growth as employees within the startups (Audretsch, Lehmann, & Warning, 2017).

2.2.4 Corporates

Corporates are often seen as the incumbent players that startups try to disrupt with a more agile and innovative approach (Freeman & Engel, 2007). However, when it comes to resources and experience, corporates have an edge over startups. Hence, various frameworks and theories

exist that try to establish a path to startup and corporate interaction. For example, Weiblen and Chesbrough (2015) describe four corporate engagement models with startups based on the dimension of equity involvement and direction of innovation flow. Firstly, corporate venturing focuses on equity investment in non-core markets to participate in the success of external innovation. Secondly, corporate incubation allows for inside non-core innovation by intrapreneurs. Thirdly, startup programs (outside-in) that have the form of incubators or accelerators enable startups to work on their ideas with the help of the organization. Lastly, startup programs (inside-out platform) make startups use corporate technology to build their products, such as app economy by Apple or Google Android.

Within the entrepreneurial ecosystem, startups and corporates spur innovation through competition while at the same time benefit from cooperation. Not only is the connection between these two actors important in the early stages of a venture, but corporates also need to play an active role within this ecosystem to identify potential M&A targets for strategic or financial purposes of the organization (Onetti, 2019).

2.2.5 Political decision-makers

Political decision-makers are interested in facilitating regional economic growth to benefit, among others, from tax income. Hence, politicians often orientate themselves at successful examples. In entrepreneurship the probably most well-known ecosystem is in Silicon Valley. However, Isenberg (2010) disagrees with the widely perception of politicians to replicate this ecosystem by stating that the Silicon Valley "ecosystem evolved under a unique set of circumstances: [...] and pure luck, among other things". Moreover, Isenberg proposes several key principals that government leaders should focus on; 1) shape the ecosystem around local conditions such as natural resources, geographic location, culture or human capital, 2) engage the private sector from the start because it has the motivation to develop profit-driven businesses, 3) favor the high potentials that address large potential markets, 4) get a big win on board to create inspiration and reduce perceived entrepreneurial barriers to entry, 5) tackle cultural change to make entrepreneurship desirable and a valid career path, 6) stress the roots and let startups find their own way to market, 7) support existing and emerging ecosystem to help them grow organically rather than creating new ones from scratch and 8) reform legal, bureaucratic and regulatory frameworks to simplify the venture formation.

With the relatively new and fast-moving entrepreneurial ecosystem, the standard tools of "business-friendly policy, such as tax incentives, grants, and local regulations, have little relevance to their success or to the vitality of local entrepreneurial ecosystems" (Auerswald,

2015, p. 13). Therefore, policy makers have to listen and communicate with the actors within the ecosystem because their main role is to provide the optimal conditions for entrepreneurs to create high-growth businesses.

2.3 Identifying and measuring entrepreneurial ecosystems

After defining entrepreneurial ecosystems and identifying its stakeholders, it is important to understand how to identify and measure the performance of the ecosystem. The possibility of accurately identifying and measuring an ecosystem's performance allows for cross-regional and cross-country comparison to identify best practices, trends and ultimately develop strategies for improvement.

The entrepreneurial ecosystem approach does not consider "traditional statistical indicators of entrepreneurship, such as self-employment or small businesses" (F. Stam & Spigel, 2016, p. 2). It instead focuses on high-growth startups or scale-ups as the source of innovation, employment, and economic growth (Mason & Brown, 2014; E. Stam et al., 2012). A quantitative study by Wong, Ho, and Autio (2005) underlines that only high growth firms, and not new firms in general, are responsible for job creation and economic growth.

Scholars provide multiple factors explaining the success of an entrepreneurial ecosystem. Widely used are the nine attributes by Feld (2012), the six distinct domains of Isenberg (2011) and the eight pillars of the World Economic Forum (2013). These three approaches largely overlap. In general, the factors encapsulate several main topics; 1) accessibility to markets, 2) human capital, 3) funding & finance, 4) cultural support, 5) regulatory framework, 6) universities as innovation catalysts and 7) other support systems. It becomes evident that the stakeholders within the system facilitate the pillars of an entrepreneurial ecosystem. Each of the actors within the ecosystem fills out one or multiple of the pillars.

Measuring how successful an ecosystem is, turns out to be a challenging undertaking. Scholars proposed different approaches depending on the definition of the entrepreneurial ecosystem and what it encompasses. Ács, Autio, and Szerb (2014) define three categories of measurement: output, attitude, and framework indicators. An example of the output measure is the Global Entrepreneurship Monitor (GEM), which records the self-employment rate within countries. The attitude measure focuses on attitudes relating to entrepreneurship, such as the preference for self-employment. However, this measure only provides insights into public opinion regarding potential entrepreneurial activity since opinions do not necessarily transform into action. The framework indicators have a regulatory approach by looking at the procedures to register a new business, which has the same caveats as the attitude measure. Henrekson and

Sanandaji (2014) used billionaires per country to measure a successful entrepreneurial ecosystem. Stangler and Bell-Masterson (2015) proposed a comprehensive approach measuring the entrepreneurial ecosystem vibrancy. Their approach combines four distinct indicators relating to density, fluidity, connectivity, and diversity. Density describes the number of new firms, share of employment and high-tech density. Fluidity defines the movement of individuals between or within regions and cities as well as high-growth firms' density. Connectivity illustrates programs and resources available to entrepreneurs. Lastly, diversity represents economic diversification, immigration and income mobility.

It is apparent that an identification and measurement of an entrepreneurial ecosystem is hard to accomplish. As shown, researchers propose various approaches that are often hard to quantify and are difficult to judge for the overall representability.

2.4 Evaluation of the status-quo

The academic community is inconsistent in defining an entrepreneurial ecosystem and its measurement, which I have shown in the literature review. Nevertheless, recurrent themes are identifiable. Although the definition may vary, the inherent meaning has become mostly similar. Moreover, a focus on productive entrepreneurship rather than business formation in general seems to be the focus of today's entrepreneurial ecosystem research. Additionally, current research primarily discusses the identification and measurement of entrepreneurial ecosystems, whereas research largely neglected the analysis of what is going on within the ecosystem. Furthermore, scholars provide insights based on personal experience and the observation of existing and well-known ecosystems, thereby risking a tautology – "entrepreneurial ecosystems are systems that produce successful entrepreneurship, and where there is a lot of successful entrepreneurship, there is apparently a good entrepreneurial ecosystem" (E. Stam & van de Ven, 2019, p. 2). Overall, many papers approach the topic only from a qualitative perspective and not a quantitative one.

Hence, deep insights into the ecosystem dynamics are important, to assess the ecosystem from a practitioner perspective more accurately. With an in-depth understanding of the innovation direction of the ecosystem, all the other actors can systematically foster relationships and promote innovation. Therefore, this Master Thesis focuses on closing this gap of ecosystem analysis with a data-driven approach through natural language processing algorithms. More specifically, I apply the approach to the cybersecurity industry with the expectation to develop an industry-independent approach that is replicable and useful.

3. Methodology

With this Master Thesis, I present a more data-driven approach for analyzing an entrepreneurial ecosystem. This approach could be the next step in entrepreneurial ecosystem research to analyze specific industries within nations and regions more precisely. The approach is dependent on machine learning algorithms such as Latent Dirichlet Allocation (LDA), Contextualized Topic Modeling (CTM), and Correlation Explanation (CorEx). I fed the algorithms with text data describing the business of a startup in a specific industry; here, the focus was on the cybersecurity industry. In order to create an industry-independent approach that is replicable, a comparison of the models was necessary. Therefore, I based the selection of the best models on objective evaluation scores and subjective interpretation. I used these models to create topics from the text data describing different domains in the cybersecurity industry. In further analysis, I looked at the prominent entrepreneurial ecosystems in Germany, Europe, Israel, and the USA. However, the process of identifying, interpreting, and using the best model contained a lot of steps which I outline in this section.

First, I sketch a definition of cybersecurity and a short introduction of the nations in focus to provide an overall context for the industry. Second, I display a general approach to the programming implementation of the Master Thesis by describing the setup, used tools and machine learning flow before diving deeper into the technicalities and processes. The deep dive starts with an explanation of the data collection as well as data cleaning and reformatting steps. Afterward, I illustrate a comprehensive explanation of the structure and reasoning of the different algorithms in use. Next, I explain the evaluation of the models to identify the final "golden model". Last, I describe the process of assigning a topic to a startup.

3.1 Industry in focus

The methodology and approach presented in this Master Thesis are supposed to be industry independent. However, to showcase the approach, I used the cybersecurity industry. In the following, I define this industry. Moreover, I outline an explanation for the choosing of Europe with a particular focus on Germany, Israel, and the USA.

3.1.1 Cybersecurity context

The term cybersecurity developed and adjusted over the last decades because of constant innovation in the software and hardware domain. Hence, Craigen, Diakun-Thibault, and Purse (2014) integrated key concepts from previous literature to create a unified definition of cybersecurity. The authors state that "Cybersecurity is the organization and collection of

resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights" (Craigen et al., 2014, p. 17). In this context, cyberspace describes a complex environment that does not exist physically but springs from the interaction of people, software, and services on the internet (Hogan & Newton, 2015). Furthermore, Von Solms and Van Niekerk (2013, p. 100) supplement this definition by specifying that, in "cybersecurity the assets that need to be protected can range from the person him/herself [..] to critical national infrastructure".

According to Zhang (2016), cybersecurity is a highly fragmented market distinguished by security types, solutions, and services. Security types include network security, cloud security, wireless security, and others. Examples of solutions are identity and access management, encryption, data loss prevention, and many more. Lastly, services focus on consulting, design and integration, risk and threat assessment, training, and education or managed security services. Currently, the increase in frequency and sophistication of cyber-attacks, the emergence of disruptive digital technologies (e.g. IoT across industry verticals, machine learning and artificial intelligence), the data-driven economy and further globalization drives the market (Grand View Research, 2020).

An example is the trend towards using multiple software-as-a-service solutions hosted in the cloud by third-party providers. This is just a single example of the increasing reliance on internet solutions which in turn provide numerous entry-points and approaches for cybercriminals to harm consumers, businesses and governments monetarily as well as personally. Due to the multiple entry-points for harm, the cybersecurity space is strongly fragmented. In the US alone, the damage from malicious cyber activities is estimated to cost the economy between $57 and $109 billion in 2016 (Council of Economic Advisors, 2018). A specific example is in the category of data breaches. A data breach potentially causes a business financial loss, affects the organization's operations and compliance as well as damages the reputation. IBM quantifies this damage at $3,86 million for a data breach. To counter these damages, government bodies introduce regulations such as the General Data Protection Regulation (GDPR) in the EU to protect its citizens' data and privacy. Ultimately, with more people and machines having access to the internet, resulting in a more crowded cyberspace, the need for protection becomes immediate. Moreover, managers start to see cyber-attacks not as an IT issue anymore but rather a Business issue because of the harmfulness to the bottom line (Columbus, 2020).

For this Master Thesis, I chose the cybersecurity industry for two reasons. First, I selected it because of its increasing importance in the organizational, economic, social, and political

domains and its manifold application in various industry verticals. Second, the fragmented nature of the industry increased the likelihood to find latent topics with machine learning algorithms.

3.1.2 National focus: Europe, Israel, and the USA

Already outlined in the literature review, scholars researched entrepreneurial ecosystems mainly in terms of identification and measurement but not so much on a per-country level to analyze the focus of the startups within that ecosystem. In this Master Thesis, I examined three nations and their locally well-established entrepreneurial ecosystems to identify differences and similarities. I chose them based on a report from Startup Blink which maps and benchmarks startup ecosystems together with government and economic development entities (StartupBlink, 2020). The nations in focus are Europe, with particular attention to Germany, Israel, and the United States of America. I outline a short explanation of their importance and selection for this Master Thesis here, whereby a more detailed analysis exists in the discussion section.

Europe is a fragmented continent with several nations that have individual entrepreneurial ecosystems. Through the European Union, 27 member states are constantly in contact by developing EU-wide regulations, policies, and incentives. For example, in July 2016, the NIS Directive was adopted, the first EU-wide legislation dedicated to cybersecurity challenges across the Union (cyberwatching.eu, 2018). Along with the General Data Protection Regulation (GDPR) of 2016, one of the strictest data protection and privacy regulation impacting EU and non-EU companies, cybersecurity has become more prominent in the EU states (Pancholi & Strobl, 2019). Germany plays an important political and economic role within the European Union, thus being of particular interest in the analysis.

Israel maintains a position in the Top 5 countries for the number of high-tech startups despite its relatively small population of 9 million. According to Start-up Nation Central (2017), the country is especially strong in cybersecurity innovation due to factors such as an overall culture of entrepreneurship and innovation, the Israel Defense Forces serving as an incubator for high-trained professionals, and specific government commitment towards the cybersecurity industry. The country's cybersecurity space has seen an ongoing five-year trend in terms of investments in the industry, peaking at 1,4$ BN in 2019 which was an increase of 35% over 2018 (Leitersdorf & Schreiber, 2020). Hence, despite being a relatively small country and being located in a tense geolocation, Israel excels in creating innovation.

Lastly, the United States of America has a worldwide track record of creating high-tech companies in the consumer and business space. Examples such as Amazon, Microsoft, Apple, Tesla, Salesforce, and Google are just an outline of their dominance in the entrepreneurial universe. Additionally, successful ecosystems such as Silicon Valley have received worldwide prestige. Lastly, the Venture Capital investments in the USA regularly account for more than half of the worldwide total amount invested and the total number of deals (KPMG, 2020). Hence, the inclusion of the USA was only natural.

3.2 General approach

3.2.1 Setup

I performed most of the preprocessing and formatting steps, as well as the implementation of the topic modeling in a Jupyter notebook running locally. However, I ran time-consuming and computational heavy tasks such as web scraping and the Contextual Topic Modeling in a Jupyter Notebook on a virtual instance on Google Cloud Platform. Jupyter Notebook has the advantage of quick experimentation and instant insights at the expense of a structured overview when running multiple models. Therefore, I connected Neptune.ai to the Jupyter Notebook. It is a separate tool to track experiments, models and to record the data exploration in parallel to running. The tool provided an easy user interface to compare different models and their outcomes. Neptune.ai kept track of the various experiments with the LDA.

Overall, the main programming language in use was Python. For the different parts of the analysis, I used various libraries such as Numpy, Pandas, Genism, NLTK, CTM, CorEx, Matplotlib, and Geopandas.

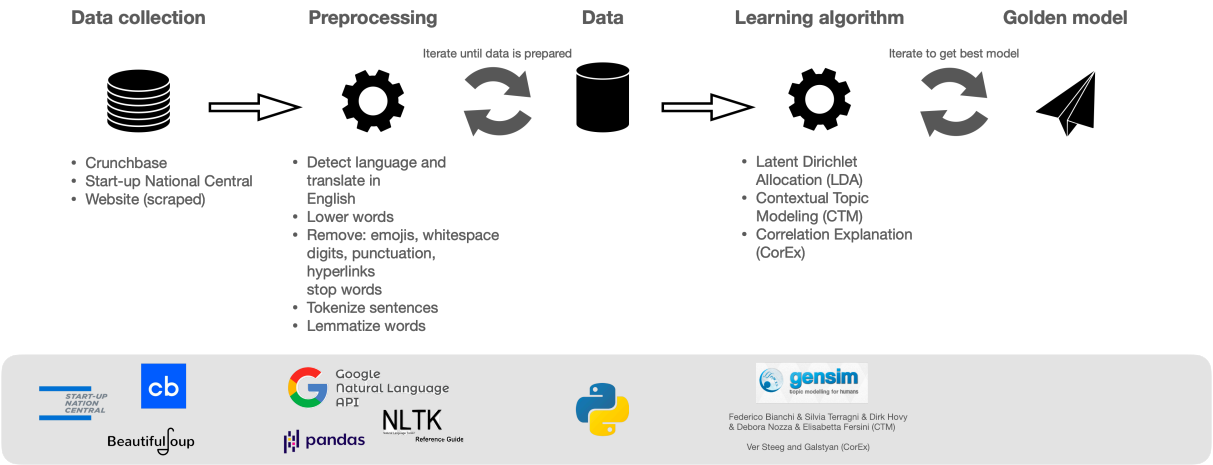
3.2.2 Machine learning flow

I used the database Crunchbase and web scraping to create a dataset. Besides, I enriched this dataset with cybersecurity companies from the Start-up Nation Central database, an Israeli NGO that fosters cross-border collaboration with businesses, governments, and other NGOs around the world. Afterward, I preprocessed and formatted the data to create the input for the models. This was an iterative process based on interim tests to ensure a final dataset with high-quality data. Then, I fed the data to unsupervised and semi-supervised machine learning algorithms. Unsupervised learning describes the process of learning without a teacher or reference which normally provides a correct answer or degree-of-error (Hastie, Tibshirani, & Friedman, 2009). Hence, in topic modeling, unsupervised learning directly infers the topics without the chance to cross-reference the results with a single point-of-truth. There is no single-

point-of truth because none of the startups is labeled with a specific topic. Semi-supervised learning topic modeling enables the injection of prior knowledge through "seed" words which encourage the model to build topics around these words (Jevtic, 2020). I used Latent Dirichlet Allocation (LDA) and Contextualized Topic Modeling (CTM) for unsupervised learning. In contrast, I used a variant of the CorEx model to set anchor words for a semi-supervised learning approach. To find the "golden model," which creates the best topics, I ran all models' multiple times on the complete dataset with several different values for the hyperparameters. The decision was based on coherence score measures and human judgment of every model. This was also an iterative process because of the vast number of variables that could be changed. I assigned every startup a topic based on the final model's topics for further downstream analysis. Figure 1 displays the complete machine learning flow of this Master Thesis and lists the tools used at every stage.

Figure 1

Description of the process to create the best machine learning model for topic modeling



3.3 Data collection

In this Master Thesis, I used multiple data sources for different purposes. These databases provided access to qualitative and quantitative data for topic modeling and further analysis, whereas interviews with industry experts supported the discussion. In this section, I describe the process of collecting the data.

3.3.1 Raw data from Crunchbase, Startup Nation Central and the startups website

I based the identification of hidden topics within a corpus of text on qualitative information. To retrieve the data for that purpose, I used Crunchbase, a database focusing on technology companies and investors. Their database contains hundreds of thousands of companies

organized in 40+ industry groups and more than 700 industries. For example, the cybersecurity industry is part of the privacy & security industry group, along with other industries such as homeland security, penetration testing, privacy, and more. With the use of Crunchbase filters, it was possible to narrow down the search. Focusing on the cybersecurity industry within the privacy & security group, along with a filter on only active companies founded after 2010, the database provided access to 5192 startups worldwide (as of September 2020). A single definition that describes how old a startup can be, to be still considered a startup, does not exist. However, for simplification, I considered only active companies founded after 2010 as startups, omitting established companies. Additionally, the database from the Israeli NGO Start-up Nation Central provided access to additional 500+ startups only from Israel.

On Crunchbase, every business provides a short description (1 – 3 sentences) and often a more detailed description (3 – 5 sentences) about their business. This description was the primary data source to identify hidden topics. However, research suggests that machine learning models dealing with text need a lot of data to create insights (Qiang, Qian, Li, Yuan, & Wu, 2020). Since the descriptions on Crunchbase were short with a focus on the general business rather than a detailed description of the technology, I assumed they have only a low explanation power. To get more detailed data, I scraped the webpage of the startup. More specifically, I scraped text labeled as "p" in the HTML source code to increase the chances of getting relevant data. In the end, the dataset for the Master Thesis contained text data on cybersecurity startups from Crunchbase, Start-up Nation Central and their respective websites.

3.3.2 Quantitative data

Crunchbase and Start-up Nation Central also provided several quantitative data points that I used for an initial understanding of the cybersecurity landscape and further analysis. For example, the founding date of the companies enabled a trend analysis of cybersecurity startups over time. Information on the companies' funding gave insights into the growth component of productive entrepreneurship.

3.3.3 Interviews

To better understand the cybersecurity ecosystem in Europe, Israel, and the United States, I conducted interviews with various stakeholders. I carefully selected the stakeholders based on their experience in the field of cybersecurity. The interviews complemented the machine learning algorithm approach and provided a more sophisticated context of the status quo and future outlooks. Moreover, I conducted interviews with data scientists that work with Natural

Language Processing (NLP) to identify the future direction of NLP and how it can affect the identification of entrepreneurial ecosystems based on text data. Table 1 shows the interview partners with certain background information.

Table 1

Overview of the interview partner

Name	Company	Position	Website
Lukas Bieringer	CISPA Helmholtz Center for Information Security	Head of Entrepreneurship & Technology Transfer	https://cispa.de/de
Oliver Spragg	Postera Capital PlanA.Earth	Advisor CTO	https://www.postera.io/ https://plana.earth/
Federico Bianchi & Silvia Terragni	Bocconi University	Researcher	Authors of CTM https://github.com/MilaNLProc/contextualized-topic-models

3.4 Data cleaning and reformatting

The biggest challenge with text data is its unstructured nature which makes sophisticated preprocessing and reformatting pivotal. In data science, the model is only as good as the quality of the input data (Gudivada, Apon, & Ding, 2017). Therefore, it was necessary to preprocess the data and to draw samples for manual checks to ensure a sufficient quality of data to proceed further.

First, I used Google's language API to identify the language of the company's description and the website text. Since the dataset contained companies from around the world, there was no guarantee that all of them provided the description and website in the English language. However, it was important to have all the information in one language to not distort and bias the topic creation. With the support of Google's API, I detected the language of the text data and translated it into English.

Next, I preprocessed the text data in several steps. Since a machine (computer) treats the same word written in lower-case and upper-case letters as different, it was necessary to bring all words in the same format. After lowering every word, the removal of unnecessary information began. In this step, I removed whitespace, emojis, digits, punctuation, hyperlinks, words with less than two letters, and stop words. Stop words are 'me', 'we', 'what', 'because', 'to', etc. In general, these do not contain meaningful information and would have distorted the topic

modeling approach. I tokenized and lemmatized the remaining words to create a common base form of the words typically written in different forms depending on the grammatical context. Last, I transformed the text data into the right format for further processing in LDA, CTM, and CorEx.

3.5 Machine-learning models

Humans are very good at understanding topics in documents based on context, experience, and intuition. Without any problem, they can follow a cooking recipe, understand the various chapters of a book and interpret the emotions conveyed in text; machines have a much harder time doing this. However, the progress in natural language processing has been tremendous in the last years. Multiple algorithms exist to infer topics from documents that would ideally match a human interpretation. Since machines can work with a larger amount of data much faster than humans, it enables the possibility to use the algorithms for the initial exploration of topics. Therefore, I used three distinct algorithms to create a model for topic generation with an unsupervised and semi-supervised approach. I fed these models with the text data from the previous step.

However, the algorithms differed in terms of implementation, speed, and inference technique. In the following, I describe each algorithm's functioning, as well as its advantages and disadvantages. Table 2 provides a short overview of the characteristics, advantages and disadvantages of the three algorithms.

Table 2

Overview of topic modeling algorithms used in Master Thesis based on the papers and own observations

	LDA	**CTM**	**CorEx**
Method	Unsupervised	Unsupervised	Semi-supervised or Unsupervised
Characteristic	• Generative probabilistic model • Bag of words • Tf-idf	• BERT • Neural variational inference	• Information theoretic framework
Advantage	• Widely used • Fast to implement • Intuitive	• Takes context into consideration	• Anchor words for model guidance • Incorporate domain knowledge

Disadvantage	• No external knowledge • No word embeddings	• Time-consuming and computationally expensive	• Need for domain knowledge
Authors	Blei et al. (2003)	Bianchi et al. (2020)	Gallagher et al. (2017)

3.5.1 Latent Dirichlet Allocation

Blei, Ng, and Jordan (2003) introduced the Latent Dirichlet Allocation (LDA). The algorithm focuses on discovering abstract topics that occur in a collection of documents, also called the corpus. LDA assumes that words carry strong semantic information and that documents discussing similar topics will use a similar group of words (Gálvez, 2017; Hu, 2009). Ultimately, the model has the notion that each document has a distribution of topics and that each topic contains a distribution of words (Figure 2). The model's inputs are a term-frequency inverse-document-frequency (TF-IDF) matrix and a list of words stemming from the preprocessing steps that focus on getting only the meaningful words. The only observable features that the model absorbs are the words that appear in the documents within the collection of documents. Other parameters are latent, such as the topic assigned to every word.

It is also important to mention that LDA is a bag-of-words model. The bag-of-words is a simple representation of text that describes the occurrences of words within a document. Since all the words are in a "bag", the algorithm ignores any information on the order or structure of the words. However, the LDA does not merely count the word occurrences within a document. It combines this approach with the TF-IDF, which normalizes the count and measures how important that term is by dividing the total number of documents (descriptions) by the number of documents that have the term. Nevertheless, the LDA does not consider any relationship between words (Y. Liu, Liu, Chua, & Sun, 2015).

Furthermore, the model does not know the number of topics in the collection of documents, but rather it is an input variable decided by the user. I carried out the process of finding the optimal number of topics by trial and error with the support of the coherence score. Trial and error comprised experimentation with the hyperparameters of the algorithm. More specifically, the variation of the threshold to create bigrams and trigrams, as well as the removal of the most frequent and infrequent words, contributed to differing coherence scores and interpretable topics. Through Neptune.ai, it was possible to track and compare all these experiments to spot the most suitable model in terms of coherence score and interpretability. I explain the term coherence score and its use in topic modeling in section 3.6.1.

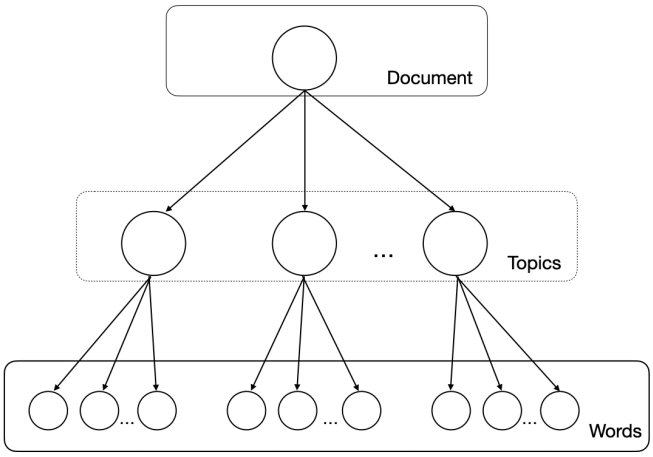
Despite its development in 2003 and hence, its relatively old age in the fast-moving computer science field, the algorithm has multiple advantages and scholars still use it in the NLP domain.

LDA works with probability distributions rather than strict word frequencies like other models. Additionally, the model enables documents to be associated with multiple topics rather than one specific topic. Another advantage is its speed compared to more sophisticated models and its intuitive implementation which does not need word embeddings nor hidden dimensions (Z. Liu, 2013).

However, there are also multiple disadvantages to LDA. The algorithm does not consider context or external knowledge because it only relies on the bag-of-words representation rather than a contextualized representation (Bianchi, Terragni, & Hovy, 2020). Moreover, it is not possible to influence the topics due to its unsupervised nature. Lastly, there is no objective metric to determine the best choice of hyperparameters that give the most accurate topics. Metrics such as coherence score and perplexity are just poor indicators for the overall quality of the model, which I discuss in more detail in the next chapter.

Figure 2

Network topology of LDA latent topics



3.5.2 Contextualized Topic Modeling

Bianchi, Terragni, and Hovy (2020) introduced a new approach to topic modeling by including contextual information; precisely the ingredient that was missing in LDA. Their study displayed a significant increase in topic coherence compared to standard LDA. With support from pre-trained Bidirectional Encoder Representations from Transformers (BERT) sentence embeddings and hence, the contribution of more contextual knowledge, the researchers were able to create more meaningful and coherent topics from their four datasets. Due to the success, the authors published their Contextualized Topic Modeling (CTM) approach in a Python

package which I used in this Master Thesis as an alternative topic modeling algorithm to the classical LDA.

Google AI team introduced BERT in 2018, which was a game-changer to the field of Natural Language Processing. The BERT language model was pre-trained with a large corpus such as articles from Wikipedia. With a combination of preprocessing steps and the complete raw text, the model can infer the language's syntax. In a second step, the model is fine-tuned with a corpus from task-specific data; here, it was the description of startups.

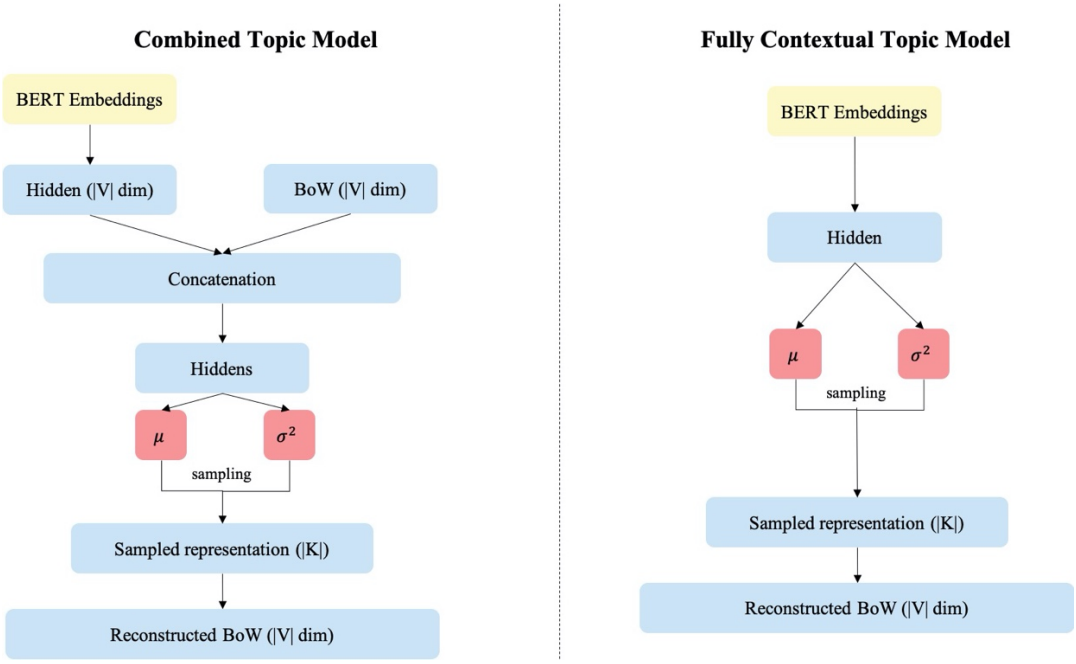
CTM provided two different methodologies. The first one combined BERT embeddings from the full-text description of the startups with a bag of words from the preprocessed description, while the second method used only BERT embeddings. Figure 3 shows the architecture of CTM which Bianchi, Terragni, and Hovy (2020) also outline in their paper.

The main advantage of the CTM is the inclusion of a pre-trained model that incorporates external knowledge in the formation of topics, which represents a significant improvement over the LDA which does not take context into account. Moreover, the pre-trained BERT model operates in a bidirectional context which is superior to previous language models that operate in a unidirectional context (Devlin, Chang, Lee, & Toutanova, 2018). This is important because the human understanding of a text is also bidirectional. When using a pre-trained model, it is possible to add one additional output layer to create new models for a specific task; in the case of CTM, it was topic modeling.

Nevertheless, CTM was more complex to implement, not as intuitive to tweak hyperparameters, and computationally heavier than the LDA. Moreover, the algorithm took a long time to run compared to the other models. These obstacles derive from the chosen BERT transformer, which in this case was the same as in the paper of Bianchi, Terragni, and Hovy (2020), namely the "*bert-base-nli-mean-tokens*". Hence, it is important to balance the sophistication of the algorithm resulting in potentially more coherent and meaningful topics with the computational intensity and time consumption. Lastly, the same flaws that existed in the LDA algorithm, such as no influence on topic creation and unsatisfactory indicators for topic evaluation, persisted.

Figure 3

High-level sketch of the Neural Topic Modeling with BERT embeddings from Bianchi et al. (2020, p. 2)



3.5.3 Correlation Explanation (CorEx) enhanced with Word2Vec

Gallagher, Reing, Kale, and Ver Steeg (2017) developed the Correlation Explanation (CorEx), the last topic modeling algorithm used in this Master Thesis. CorEx is rooted in the concept of total correlation, an information-theoretic measure that allows the model to learn "maximally informed topics" (Gallagher et al., 2017). In this context, the authors do not understand the term correlation in its classical sense but rather as dependence. Contrary to the LDA, the CorEx model relies on the concepts of entropy and mutual information to describe dependence through total correlation to infer topics. Moreover, CorEx works without additional assumptions, whereas LDA requires the specification of hyperparameters to characterize the generative process (Jaycocks, 2019).

The previous algorithms utilized unsupervised learning. CorEx provided the opportunity for another unsupervised learning approach but also a semi-supervised method. Through semi-supervised learning, it was possible to set anchor words. These anchors represented words associated with a topic. The setting of anchor words makes sense if specific terms relating to a topic are known a priori and when a specific topic is of interest but only present in a small subset of the documents (Jaycocks, 2019). The small subset of documents receives a topic in a binary manner, rather than a probability distribution as in the case of the LDA. I administered two steps to create the anchor words. First, with thorough research in the cybersecurity field, I

applied domain knowledge. Second, with a Word2Vec model, a neural network to learn word embeddings, I chose similar words from within the corpus (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

Within the scope of this Master Thesis, I trained the Word2Vec model on the dataset mentioned above, contrary to the alternative that uses a pre-trained model. Word2Vec presents the words in their written way in the corpus, which is superior to the assumptions made with enabling domain knowledge. For example, based on domain knowledge, I believed the word 'multifactor authentication' to be present within a topic. This could well be true, but the text description used the short version "mfa" which has the same meaning. Figure 4 shows the ten most similar words to the term "authentication" based on the Word2Vec embeddings. The score shows the cosine similarity between the words, whereby a 1 would indicate the exact same word, -1 an opposite word and 0 independency. I created a rule of thumb through trial and error that a score above 0,68 provides interesting words. Hence, I added all the words with a score above 0,68 to the set of anchor words. Furthermore, selecting the set of anchor words was reiterative because of the possibility to vary hyperparameters and the need to add or remove words depending on the outcome of the model. Enhancing the anchored CorEx model with most similar words from a Word2Vec model is a new approach that tries to automate and accelerate the selection of insightful anchor words.

According to Gallagher et al. (2017), CorEx made it possible to define a weight variable for the anchor strength, representing the model's intervention. A high weight pushed words stronger into a certain topic, whereas a low weight resulted in more flexibility of the word assignment. Ultimately, an increased confidence that a specific topic should exist within the dataset could be forced through a high weight. This meant that despite the existence of a low number of words to make it a standalone topic, CorEx channeled the model in a specific direction. Here, I followed the authors' guidelines by setting the anchor strength at 4, nudging the topic more strongly in a specific direction.

Figure 4

Example of Word2Vec most similar words which inspire anchor words

```
1  w2v_model.wv.most_similar(positive='authentication', topn=10)
executed in 10ms, finished 10:26:31 2020-11-10
[('passwordless', 0.7742965221405029),
 ('multifactor authentication', 0.7567595839500427),
 ('authorization', 0.7340003252029419),
 ('biometrics', 0.7005100846290588),
 ('mfa', 0.689814567565918),
 ('two factor', 0.6728553771972656),
 ('biometric', 0.6642657518386841),
 ('login', 0.6498785614967346),
 ('authomate', 0.6463791728019714),
 ('identity', 0.6095454692840576)]
```

3.6 Measurement and understanding

After running several models with different hyperparameters, it was necessary to choose the "golden model" that created the most coherent and interpretable topics. However, the pick of the best model was not straightforward. Therefore, I combined objectively computed metrics and human judgment to find that model.

3.6.1 Coherence scores

The topic modeling algorithms described above have different approaches to creating topics. Hence, it was necessary to compare them to find the best model that creates interpretable and coherent topics. According to Röder, Both, and Hinneburg (2015), topics are said to be coherent if the words support each other. For example, a set of words comprising [virus, malware, ransomware] is more coherent than a set of words such as [soccer, cyber, pizza]. Research still largely discusses how to quantify the coherence of such a set so that it reflects the opinion of a human (Rosner, Hinneburg, Röder, Nettling, & Both, 2014). Since humans are known to have different opinions, it complicates the development of an objective coherence measure. The collection of some words in a topic may be understandable by one human but not by another.

Multiple coherence measures exist that differ in their way of calculation. Nevertheless, to evaluate the performance of the models, I used three coherence scores, namely C_v, C_npmi, and C_umaas.

Still, researchers point out that metrics such as coherence scores are just a supporting indicator for the topics' overall quality (Röder et al., 2015). A high coherence score does not necessarily mean that it resembles the interpretation of a human. Therefore, I used the coherence score as an initial starting point to assess the model's overall quality. It provided a direction for the selection of the number of topics in the model.

3.6.2 Human judgment

I used human judgment to equilibrate the flaws of the coherence measurement mentioned in the previous paragraph. After the topic creation, I manually inspected the topics to understand the theme of the words within the topics and to assess their meaningfulness. Depending on the algorithm, several topics might have contained the same words. Although the words were in a different order creating a different coherence score, it did not make sense to have such a scenario. The goal was to create standalone and distinguishable topics. Hence, the manual inspection confirmed the meaningfulness of the created topics. Moreover, the created topics contained set of words which I assigned a specific topic name. For example, the set of words – *fraud, payment, transaction, trading, kyc, financial, credit card, banking, card, fraud prevention* – was interpreted as financial fraud.

3.7 Topic assignment and further analysis

Due to the outcomes of the models presented in the next section (4. Results), I chose the CorEx model to assign the final topics to the startups for further analysis. CorEx assigns in this step a single, multiple, or no topics to a startup. Hence, if the model could not infer a topic based on the startup description, no topic was assigned, which in turn required a manual assessment. Now, a topic comprised a set of words in which each word had a mutual information score. With Python, I matched each startup description with all possible topics to identify the sum of the correlation. Normally, the scores are around 0,005 to 0,1 which makes a comparison and intuitive understanding difficult. Therefore, these scores were normalized to make them comparable and to put the values between 0 and 1, making an understanding more natural (Lakshmanan, 2019). Afterward, I defined a threshold score enabling the assignment of the topics. Based on manual inspection, I set the threshold at 0,5. If the match between topic and description was above 0,5, I assigned that topic to that specific startup. On the other hand, if it was below 0,5, I did not assign the topic. This technique made the allocation of multiple topics to a single startup possible. In cybersecurity, such a scenario makes sense because a company is likely to describe multiple parts of its business in its description. For example, a startup describing its business as an authentication tool based on blockchain technology should have topics such as *Blockchain* and *Authentication security.*

4. Results

The previous section objectively described the methodological approach to the Master Thesis. Three algorithms were in the focus, namely LDA, CTM, and CorEx which differed in terms of implementation and the process of inferring topics.

This section aims to describe the results of these algorithms, with the first part discussing the descriptive statistics that provide a better understanding of the dataset. Second, I compare the models to showcase the different approaches' outcomes and outline the decision-making on the best model. The comparison goes hand in hand with a display of the topics that I created through the algorithms. Fourth, I illustrate the topic distribution per country. The section concludes with a portrayal of the entrepreneurial hotspots on a regional level.

4.1 Descriptive statistics

In this part, I provide a short description of the utilization of the databases as well as the final dataset. This enables the reader to better understand the magnitude of available data, especially on a per-region basis. Moreover, I display the founding of companies per year in each region to grasp the differences in the registration of cybersecurity startups on Crunchbase and Startup National Central.

4.1.1 Dataset

I developed the dataset with two databases that I filtered for cybersecurity startups founded after 2010. Crunchbase provided a total of 5192 companies, whereas Startup National Central added another 508. After dropping all the duplicates and manually deleting companies that did not deal with cybersecurity or did not have a long enough description, the final dataset contained 4894 startups. Europe, Israel, and the United States represented 64,5% of these startups with an absolute count of 709, 476, and 1969, respectively. It is also noteworthy that Germany, with 102, contributes the second greatest number of startups in Europe, right behind France (103) and before the Netherlands (100). An overview of every country and its number of startups is in Appendix A.

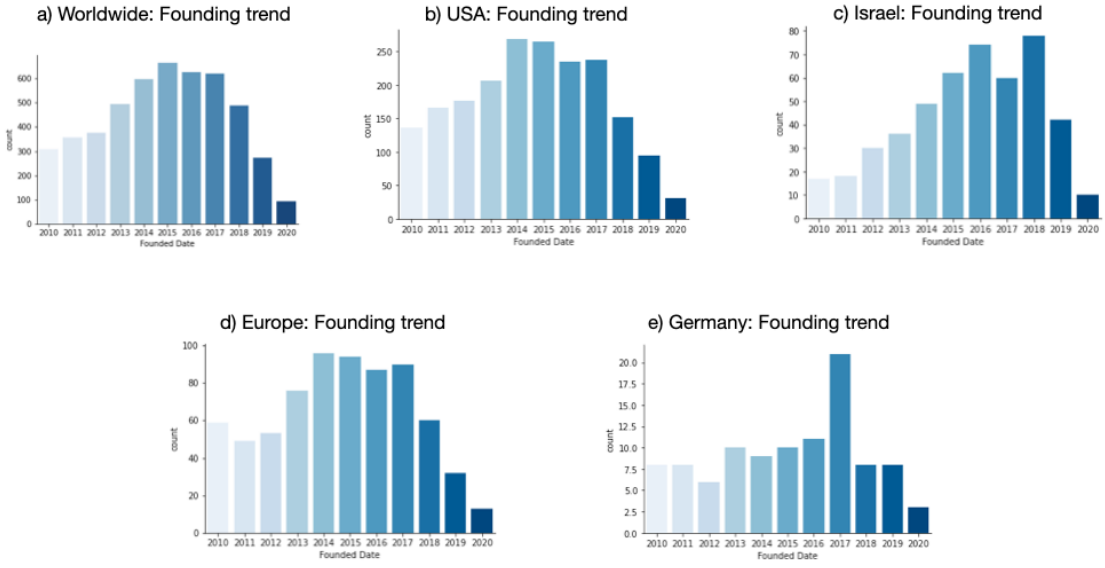
4.1.2 Founding trend of cybersecurity startups

The dataset contained information on the founding date of the startups. From Figure 5a, it becomes apparent that a continuous increase in the founding of cybersecurity startups occurred since 2010, spiking in 2015, whereas the years 2019 and 2020 show a small number of startup registration. This makes sense because Crunchbase mostly tracks companies that obtained

funding or already entered the market, which is an unlikely scenario for 12 – 18 months old startups. Consequently, although a startup is registered within a country, it does not necessarily provide its information on Crunchbase in the same year. For example, startups founded in 2020 are possibly added the earliest in 2021 which delays an accurate count of the year 2020.
Nevertheless, it is interesting to see that the distribution of founding in the United States and Germany is similar whereas the distribution in Israel is slightly lagging by one to two years. Additionally, the spike in Germany in 2017 is eye-catching, showing a 100% increase in startup founding compared to the previous year.

Figure 5

Startups founded per year since 2010 in Europe, Germany, Israel, USA, and worldwide



4.2 Model comparison

In the following part, I describe the results of the models from LDA, CTM, and CorEx. I segmented this description into three parts. The first part deals with the experiments' setup, which is responsible for the inference of different topics. Next, I compare the results of the experiments with a specific focus on the various coherence measures. Last, I display the actual set of words constituting the final topics of the respective model.

4.2.1 LDA

The meaningfulness and interpretability of the LDA topics depend highly on the number of unique tokens in the dictionary of the model. In other words, the dictionary contains all the possible words that can be assigned to the topics. In order to understand the influence of this

factor on the coherence score and human interpretability, I ran multiple experiments. Tables 3 and 4 show the setup of these experiments. The columns *In at least X documents* and *In no more than X documents* show the values for parameters that filtered out tokens. For example, experiment 2 describes a dictionary that does not have words that were not in at least five documents and does not have words that were present in over 80% of the documents, resulting in a dictionary containing 4031 words. Moreover, the LDA needs a parameter indicating the number of topics. Here, I executed the model multiple times to calculate the coherence score on a range of topics from 5 to 50, in steps of three, that is 5, 8, 11, ..., 47.

Through experimentation, it was possible to compare the different coherence scores to judge the performance of the model. Figure 6a) shows the results of the LDA which only included the description of the startups. It shows that the spike in coherence score is relatively early in all experiments between 5 and 14 topics. In this situation, the 4th experiment exhibits the highest coherence score of 0,475 with 8 topics. Table 5 displays the created topics from the model based on the specifications of the 4th experiment. It becomes apparent that none of the 8 topics are coherent. Although most topics clearly include words relating to cybersecurity, they are not associable with one specific domain. For example, topic 3 has a tendency towards *risk assessment*; however, other words such as *identity access, consulting firm* and *mobile apps* distort that picture.

The LDA results on the scraped website data (without Crunchbase data) are more constant within the experiments (see Figure 6b). This means that the coherence score does not vary as much as in the previous LDA. Nevertheless, the 1st experiment with the largest dictionary exhibits the highest coherence score (0,500) at 11 topics. Table 6 shows the topics created in this experiment. Some topics are more coherent than the topics from the "Description LDA", such as 7, 9, and 11, clearly dealing with *Identity & access management, Vulnerability testing,* and *Blockchain & transactions*. Other topics have a wild mix of words and topic #2 does not make sense in the context of cybersecurity.

Based on these results, I concluded that neither the LDA on the startup description nor the LDA on the scraped data from websites created insightful and coherent topics. However, as already indicated in the methodology, the coherence measures do not perfectly resemble the human interpretation, justifying these inconclusive results. Although the topics were not entirely coherent, they provided a good direction for further experiments with other algorithms. Lastly, it is necessary to mention that I only scraped 3762 websites. I did not scrape the other 1132 due to missing permission of the website owners. Hence, the LDA on the scraped website had more absolute data to work with but did not resemble all potential businesses because of that

limitation. Thus, I fed the following models (CTM and CorEx) only with the complete dataset from Crunchbase, excluding the information extracted from the websites.

Table 3

LDA experiments on the descriptions of the startups

Experiment	In at least X documents	In no more than X documents (%)	Tokens
1	1	1,0	19002
2	5	0,8	4031
3	10	0,6	2704
4	20	0,8	1569

Table 4

LDA experiments on the website text of the startups

Experiment	In at least X documents	In no more than X documents (%)	Tokens
1	1	1,0	44946
2	5	0,8	9186
3	10	0,6	6485
4	20	0,8	3995

Figure 6

Coherence score based on a) description and b) website

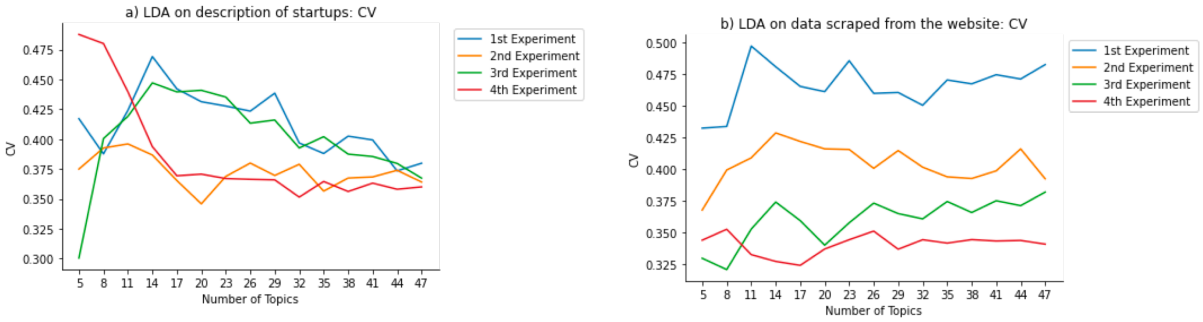


Table 5

Topics created from LDA based on startup description

#	Topic
1	big_data, company_specializes, social_medium, detection_response, supply_chain, across_globe, personal_data, social_network, patented_technology, operation_center
2	incident_response, also_offer, public_cloud, best_practice, data_science, smart_contract, vulnerability_management, fraud_prevention, visibility_control, global_leader
3	risk_management, risk, identity_access, consulting_firm, enables_user, attack_surface, mitigate_risk, mobile_apps, fully_automated, reduce_risk
4	artificial_intelligence, software_development, easy_use, mobile_app, year_experience, regulatory_compliance, without_need, secure_communication, mission_help, personal_information
5	penetration_testing, data_protection, operating_system, data_loss, data_recovery, sensitive_information, intellectual_property, enables_organization, fraud_detection, vulnerability_assessment
6	network, cloud, information, secure, platform, device, hardware_software, system, enterprise, connected_device
7	machine_learning, threat, attack, platform, network, cloud_storage, law_enforcement, enterprise_government, anomaly_detection, information_visit
8	threat_intelligence, insider_threat, government_agency, digital_transformation, founded_based, financial_institution, web_mobile, '_mission, secure_compliant, silicon_valley

Table 6

Topics created from LDA based on startup website

#	Topic
1	information, email, website, privacy, online, site, people, may, process, web
2	republic, state, island, democratic, com, people, arab, und, der
3	system, industry, project, government, innovation, consulting, startup, digital, design, research
4	**cloud**, **platform**, risk, **network**, compliance, **infrastructure**, enterprise, secure, manage, control
5	vpn, internet, network, wifi, quantum, proxy, speed, private, ddos, country
6	system, network, recovery, camera, home, drive, computer, data_recovery, control, video
7	**device**, **access**, **secure**, file, mobile, **identity**, **password**, iot, **authentication**, **encryption**
8	privilege, web, twittercomiwebstatus, cloudflare, identity, account
9	**threat**, **attack**, **vulnerability**, network, system, risk, **detection**, intelligence, analysis, **protection**
10	website, certificate, ssl, ssl_certificate, necessary, function, personal, website_function, site, includes

11	**blockchain**, **payment**, **transaction**, **wallet**, **crypto**, **asset**, **exchange**, **contract**, **digital**, **cryptocurrency**

4.2.2 CTM

The results of the LDA were not very insightful and interpretable. As already mentioned in the methodology section, a more sophisticated approach that also includes the context of the words was necessary. Table 7 displays the setup of this approach. For this algorithm, the main distinction was the methodology, contextual or combined, and the number of topics. Overall, it becomes apparent that they both outperform the LDA when looking at the coherence score CV. In the LDA, the highest score was around 0,500, whereas CTM produced a score of 0,560 which is an increase of 12%. Additionally, I observed that the highest coherence score is associated with a different number of topics. In the LDA it was 11, while the CTM spikes at 17. Moreover, the values of the NPMI and UMASS scores are also close to 0 compared to the other experiments. Hence, I used the contextual methodology with 17 topics to create topics.

Table 8 shows these 17 topics. I allocated Topics 1, 3, 6, 10, 12, and 14 to the domains *threat intelligence, early-stage startups, services, authentication, threat intelligence,* and *data privacy*, respectively. In general, the majority of the topics contain words relating to cybersecurity and innovation. Only a handful of words, such as *beijing, costly, and headquartered,* do not provide value. Nevertheless, the words within the topics are often associated with separate domains and do not inherently describe one specific cybersecurity field. For example, topic 3 deals with early-stage startups in general, but there is no reference to cybersecurity. The high rank-biased overlap (RBO) in Table 7 indicates how diverse the topics are from each other in terms of different words and the order of the words. Although the scores are high (above 0,900), several words exist in multiple topics. These overlaps are especially apparent in topics 1 and 12, as well as in 2 and 17. Words such as *intelligence, threat, risk, data, privacy, consulting, authentication,* and *password* also exist in the topics created by the LDA.

In direct comparison with LDA, CTM objectively creates better topics based on the evaluation metrics. However, from a subjective perspective evaluating the topics with human capabilities, both algorithms do not create standalone topics that are meaningful and interpretable. Nevertheless, the algorithms provide a foundation and direction for further exploration.

Table 7

Scores of the combined and contextual methodologies of the CTM algorithm

Combined CTM				
# of topics	**CV**	**NPMI**	**UMASS**	**RBO**
8	0,531	-0,130	-7,777	0,928
11	0,553	-0,132	-7,940	0,910
14	0,551	-0,096	-6,439	0,935
17	0,532	-0,092	-6,268	0,925
20	0,548	-0,068	-5,780	0,926

Contextual CTM				
# of topics	**CV**	**NPMI**	**UMASS**	**RBO**
8	0,557	-0,129	-8,179	0,944
11	0,540	-0,095	-6,345	0,930
14	0,524	-0,063	-5,587	0,944
17	**0,560**	**-0,052**	**-5,441**	**0,951**
20	0,544	-0,095	6,423	0,937

Table 8

Topics created with the CTM

#	**Topic**
1	cyber, **security**, companies, **intelligence**, **threats**, **risk**, cybersecurity, **threat**, organizations, **attacks**
2	cloud, solutions, network, enterprise, services, software, provider, networks, service, mobile
3	**startups**, **early**, **entrepreneurs**, career, **accelerator**, **stage**, women, tech, **seed**, **emerging**
4	attacks, threats, solution, protection, protect, fraud, time, advanced, detection, mobile
5	world, technology, experience, team, new, best, development, digital, great, years
6	**services**, **consulting**, **business**, **firm**, **information**, **solutions**, team, **implementation**, **support**, **experience**
7	unmatched, footprint, seeks, compared, granular, breakthrough, values, beijing, costly, helped
8	compared, unmatched, initial, costly, footprint, values, helped, granular, dynamically, seeks
9	platform, time, ai, real, threat, intelligence, solution, based, application, detection
10	**secure**, **users**, **login**, **password**, use, share, **passwords**, user, never, **authentication**
11	online, website, video, businesses, site, users, free, social, videos, know
12	**intelligence**, **threats**, **threat**, **cyber**, time, **attacks**, **advanced**, **real**, **attackers**, platform
13	unmatched, beijing, suspicious, footprint, compared, dynamically, costly, scanner,

granular, values

14	**data**, **privacy**, **secure**, **access**, **encryption**, solution, **keys**, without, devices, **authentication**
15	founded, company, software, cybersecurity, solutions, services, provides, offers, information, headquartered
16	security, cyber, team, experience, cybersecurity, services, intelligence, companies, information, solutions
17	cloud, data, platform, security, enterprise, compliance, access, applications, management, organizations

4.2.3 CorEx

The results of the LDA and CTM did not provide standalone topics for further processing but rather an indication of further direction. With the insights from these models, I used the semi-supervised method from CorEx. Through a reiterative process of setting anchor words based on domain knowledge, results from LDA and CTM, as well as closely related words revealed by Word2Vec, I created 17 topics with CorEx. The reiterative process was 10 cycles long. In these cycles, I adjusted and varied the anchor words, the total number of topics, and the vocabulary length to optimize the outcome. In this scenario, I utilized Word2Vec to enrich the anchor words created a priori and to identify other words, synonyms, or related terms to the set of words of the anchor.

Table 9 displays the set of words associated with each topic. Based on these words, I manually selected a final domain name. It is apparent that the topics are widely diverse and that almost none of the words within a topic is meaningless. Compared to the previous algorithms, the semi-supervised approach with CorEx creates standalone interpretable topics and makes sense from an objective and subjective perspective. None of the words exist within multiple topics and not a single topic is attributable to multiple domains.

I divided the 17 topics further into four categories: Focus Area, Technology, Field of Application, and Service. The first category, **Focus Area,** describes topics that deal with a specific domain of cybersecurity. These are *Testing & assessment, Identity & access management, Application security, Fraud & transaction, Network & infrastructure security, Operational technology security, Data privacy & data protection, Messaging security, Cloud security, Forensics* and *Web security*. The **Technology** category defines a technology in use and only consists of the topic *Blockchain*. The third category, **Field of Application**, characterizes specific industries highly dependent on cybersecurity solutions such as the *Internet of things, Autonomous vehicles* and *Smart homes*. Lastly, the topics *Consulting* express

a **Service** rather than a product or technology. Companies focusing on *Education* are assignable to the category of **Focus Area** as well as **Service**.

Table 9

Topics created with CorEx semi-supervised machine learning algorithm

	Topic
Testing & assessment	testing, penetration, penetration testing, vulnerability, assessment, vulnerability assessment, ethical hacking, penetration test, pen, red teaming, security testing, bug, bounty, bug bounty, vulnerability scanning
Identity & access management	authentication, identity, password, biometric, passwordless, identity access, biometrics, authorization, multifactor authentication, biometric authentication, mfa, two factor, access management, identity access management, digital identity
Blockchain	blockchain, decentralized, smart contract, protocol, blockchain security, distributed ledger, ethereum, computation, blockchain technology, blockchainbased, contract, ledger, decentralized application, distributed, cryptocurrency
Internet of things	iot, iot device, connected, iot security, connected device, internet thing, device, connectivity, thing, thing iot, internet thing iot, secure iot, medical device, iot security solution, edge
Application security	application security, application, serverless, ddos, application security solution, serverless security, web application security, web application, distributed denial service, distributed denial, ddos attack, application security company, service ddos, denial service ddos, denial service
Autonomous vehicles	drone, vehicle, autonomous, car, unmanned, autonomous vehicle, aviation, cybersecurity protection, automobile, fleet, los, notch, aircraft, truck, gps
Consulting	consulting, consultancy, advisory, consulting service, staffing, security consulting, consulting company, consulting firm, security consultancy, security consulting service, information security consulting, technology consulting, cybersecurity consulting, cybersecurity consultancy, firm
Education	training, education, educate, gamified, security awareness, awareness, training platform, security training, security awareness training, cyber security training, security education, skill, cybersecurity awareness, teach, training company
Fraud & transaction	fraud, payment, transaction, trading, kyc, financial, credit card, banking, card, fraud prevention, fraud detection, credit, authenticity, bank, fraudulent

Smart home	home, smart home, connected home, home security, home network, smart, security home, alarm, home business, homeowner, device home, home office, router, protect connected, home automation
Network & infrastructure security	network security, network, network security company, network security service, computer network security, computer network, network security solution, security company located, company located, security company, security network security, computer, offer network, located, network security monitoring
Operational technology security	industrial, industrial control, control system, industrial control system, scada, automotive, critical infrastructure, industrial cybersecurity, control, industrial iot, system, industrial network, industrial automation, critical, manufacturing
Data privacy & data protection	privacy, data protection, gdpr, regulation, privacy regulation, ccpa, hipaa, data security, data, data privacy, pci, protection, privacy compliance, psd, gdpr ccpa
Messaging security	email, file, email security, email address, secure email, email service, message, encrypted, messaging, file sharing, confidential, instant messaging, confidential data, chat, business email
Cloud security	cloud, cloud security, public cloud, aws, cloud service, azure, cloud infrastructure, cloud solution, enterprise cloud, cloud service provider, cloud security service, provides cloud, aws azure, cloud computing, aws cloud
Forensics	investigation, incident response, forensics, threat hunting, digital forensics, forensic, response, incident, hunting, ediscovery, threat, orchestration, security operation, detection, detection response
Web security	web, web security, ssl, hosting, certificate, ssl certificate, web hosting, domain name, domain, hosting solution, vpn, geotrust, comodo, vpn review, name

To sum up, the main advantage of CorEx is its option to include domain knowledge in the formation of topics, which was not possible within the LDA and CTM analysis.
A disadvantage that becomes apparent when using anchor words is the necessity to have sufficient domain knowledge. It involves slightly more work than the other models due to the semi-supervised reiterative nature of the model. Lastly, the same flaws as the previous models include the objective measurement of the model's performance and the self-chosen number of topics.

4.3 Topic assignment

I created a total of 17 topics with the semi-supervised method of CorEx. Based on the cutoff level (0,5) described in the methodology, 3596 startups have only one topic, 1063 have two

topics and 49 have three topics. This is possible because of the four categories defined in the previous section. The assignment method leaves room for multiple assignments, which is realistic since a business can utilize a technology such as *Blockchain* within a specific focus area such as *Identity & access management*.

Table 10 shows the topic distribution in Germany, Europe, Israel, the USA, and worldwide. Here, I show the distribution in relative terms to make it comparable; otherwise the USA representing almost 45% of the dataset, would completely outweigh the other nations. For example, the 11,76% for *Testing & assessment* in Germany describes the share of startups in Germany that have this topic assigned to them. More specifically, 12 out of the 102 startups in Germany deal with the Focus Area of *Testing & assessment* in their business.

Overall, the topics *Testing & assessment, Identity & access management, Network & infrastructure security, Data privacy & data protection* and *Cloud security* are represented to a large extent indicated by a value above 10%. Other topics are in the minority, such as *Application security, Autonomous vehicles, Education, Smart home, Operational technology security, Messaging security* and *Web security.*

In addition to the general distribution of the topics across startups, there are several differences across the countries. Some of the topics are almost evenly distributed across these geographical areas, such as *Identity & access management, Application security, Education, Financial fraud, Operational technology security, Data privacy & data protection, Messaging security,* and *Web security*. Other topics, on the other hand, are predominately present or underrepresented in a specific location. The topic *Testing & assessment* is below the worldwide value in all locations, especially in Israel, where it shows an almost 4% difference to the worldwide value. The *Blockchain* technology is more often used in Germany than in Israel or the USA. Israel is the pioneer within the *Internet of things* domain, as indicated by the relatively high share of 10,71%. Germany spearheads within the *Autonomous vehicle* realm with a share of 3,92% and *Smart home* with a share of 4,90%; however, in absolute terms, this entails only four and five startups, respectively. The service *Consulting* prevails in Europe as well as the USA, whereas Israel and Germany have less than 6% and 4%, respectively, of their companies focusing on this domain. *Network & infrastructure security* appears most often in Europe and Israel. The largest gaps exist within the topics of *Cloud security* and *Forensics*. With more than 21%, every fifth startup in the USA shows a relation to cloud security, while Israel, Europe and Germany lack behind with less than 15%, respectively. I observed the same distribution within *Forensics* in which the USA has a share of 8,89% while Germany shows a low share of 1,96%.

Table 10

Topic distribution in %

	Germany	Europe	Israel	USA	World-wide
Testing & assessment	11.76	12.27	8.82	11.02	12.63
Identity & access management	9.80	8.32	10.50	10.06	9.07
Blockchain	5.88	4.09	4.20	4.16	4.11
Internet of things	6.86	6.77	10.71	7.26	6.85
Application Security	2.94	1.97	2.52	3.15	2.70
Autonomous vehicles	3.92	1.55	2.10	1.47	1.68
Consulting	5.88	10.86	3.57	8.13	9.52
Education	1.96	3.24	2.94	3.86	3.66
Financial fraud	6.86	7.19	7.35	5.59	6.31
Smart home	4.90	2.12	2.31	2.03	2.15
Network & infrastructure security	12.75	16.22	15.76	13.20	14.26
Operational technology security	3.92	4.51	4.20	2.29	3.00
Data privacy & data protection	14.71	12.98	16.39	14.02	13.14
Messaging security	3.92	1.97	3.57	3.30	3.39
Cloud security	12.75	14.39	14.29	21.18	16.49
Forensics	1.96	5.50	7.56	8.89	6.99
Web security	4.90	3.10	4.20	3.45	4.11

Table 10 displays the relative distribution of the topic in the geographical areas, which provides insights into the overall focus of the startups within that region. However, the Table needs more context in absolute terms to make it comparable because Germany exhibits only 102 startups in the dataset, whereas the USA produces almost 2000. Figures 7 – 10 display the absolute distribution of the topics in every Bundesland in Germany, in the top 10 countries in the EU,

the top 10 cities in Israel and the top 10 states in the USA. The top 10 describes the countries, cities or states with the highest numbers of total startups within that geography.

Figure 7 shows the topic distribution of every Bundesland in Germany. Here, it becomes apparent that Bremen, Mecklenburg-Vorpommern, Sachsen-Anhalt, and Saarland do not have cybersecurity startups represented on Crunchbase. I divided the other Bundesländer into three groups. The first group contains Bayern and Berlin, representing almost 50% of the cybersecurity startups in Germany. Moreover, they exhibit startups from almost every topic in their area except for *Forensics* and *Autonomous vehicles* missing in Bayern and Berlin, and *Operational technological security, Smart home, Education*, and *Consulting* missing only in Bayern. The second group contains Nordrhein-Westfalen, Hessen, Hamburg and Baden-Württemberg hosting startups from roughly eight topics. Mostly noticeable is Nordrhein-Westfalen with four startups in the *Testing & assessment* field which is the highest cumulation in Germany. The third group contains the rest of the Bundesländer which only display a focus of one to four topics. Overall, the topics that are most represented in Germany deal with *Cloud Security, Data privacy & protection*, and *Network & infrastructure security*. Nevertheless, compared to the absolute numbers of these topics in Israel (Figure 9) and the USA (Figure 10), Germany seems to play only a minor role in the output of cybersecurity startups.

Figure 8 shows the topic distribution in the top 10 countries in the European Union based on the total number of startups. France, the Netherlands, Germany, and Spain are riding ahead with almost 100% coverage of all the topics in their area. Especially the topics *Cloud security, Data privacy & data protection, Network & infrastructure security*, and *Testing & assessment* are the most prominent in these countries. In the Netherlands, it is noticeable that *Consulting* services play an important part compared to the other three countries, with a 100% difference in the number of startups focusing on that topic. The other countries, such as Sweden, Ireland, Italy, Finland, Denmark, and Estonia, have evenly distributed mix of topics.

Figure 9 shows the topic distribution within the cities of Israel. However, since most cities' driving distance is roughly between one to two hours, readers should interpret the figure as a whole. Moreover, I added Tel Aviv separately because of its dominance in the country which skewed the figure's scaling. Especially three topics are prominent in Israel, namely *Cloud security, Data privacy & data protection* and *Network & infrastructure security*. Other topics of interest are *Blockchain, Identity & access management*, and *Testing & assessment*. This distribution almost mirrors the distribution in Germany and the European Union except for a larger share in Blockchain.

Figure 10 shows the topic distribution of the top 10 states with the highest number of total startups in the United States. I added California with a separate axis because of its dominance. The main focus in California is on *Cloud security*, with almost 25% of the total number of startups in that state. Additionally, the topics *Data privacy & data protection, Network & infrastructure security* as well as *Identity & access management* play a vital role in that state. The other states have mostly a mix of all the topics, although some specializations are identifiable. For example, New York shows a focus on the *Blockchain technology*, whereas Virginia exhibits a large portion of *Forensics*.

Overall, in all four geographic arrangements, a trend towards certain topics becomes apparent. These topics are *Data privacy & data protection* in Germany; *Network & infrastructure security* and *Data privacy & data protection* in the European Union; *Cloud security, Data privacy & data protection* and *Network & infrastructure security* in Israel as well as in the USA.

Figure 7

Topic distribution per Bundesland in Germany

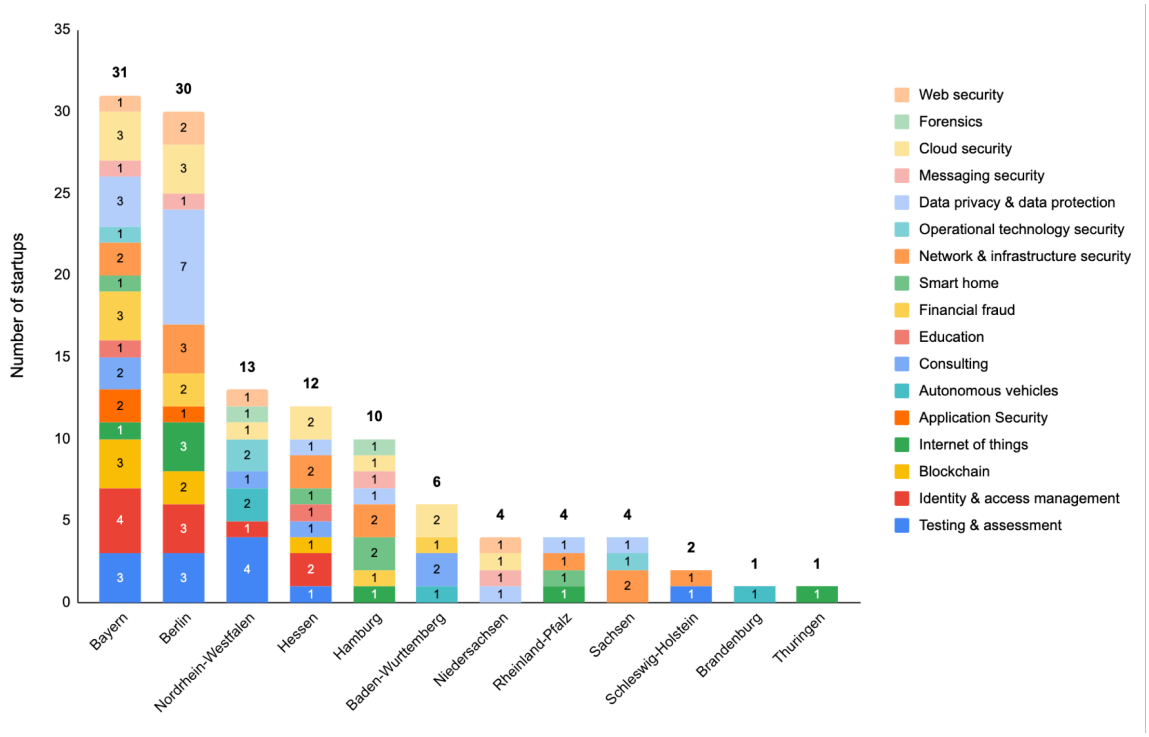


Figure 8

Topic distribution of the Top 10 countries in the EU

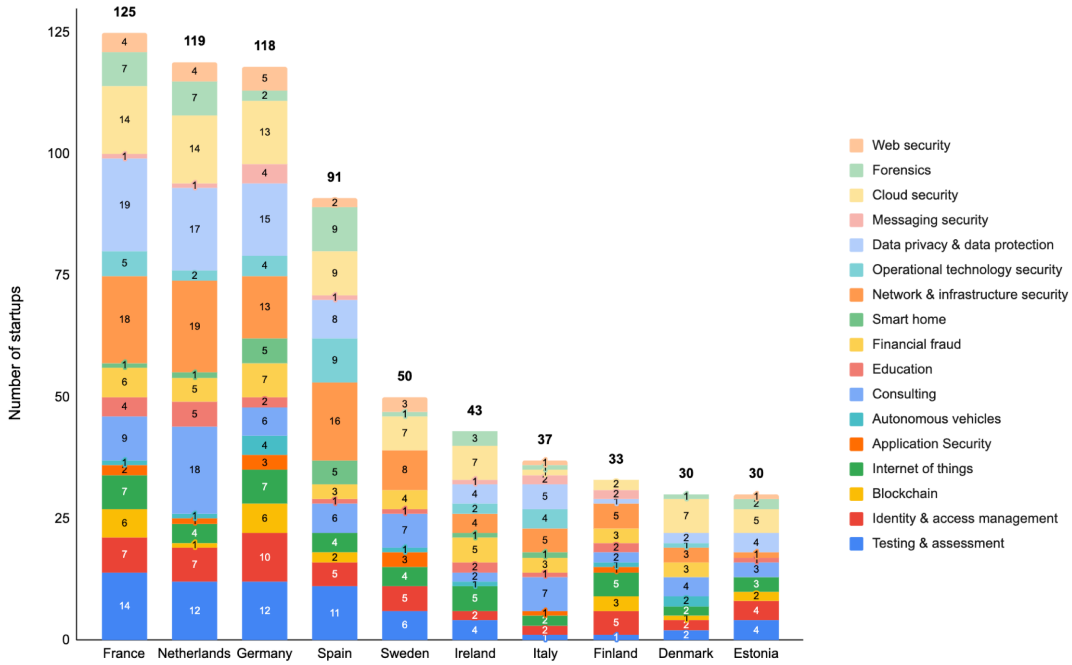


Figure 9

Topic distribution of the Top 10 cities in Israel

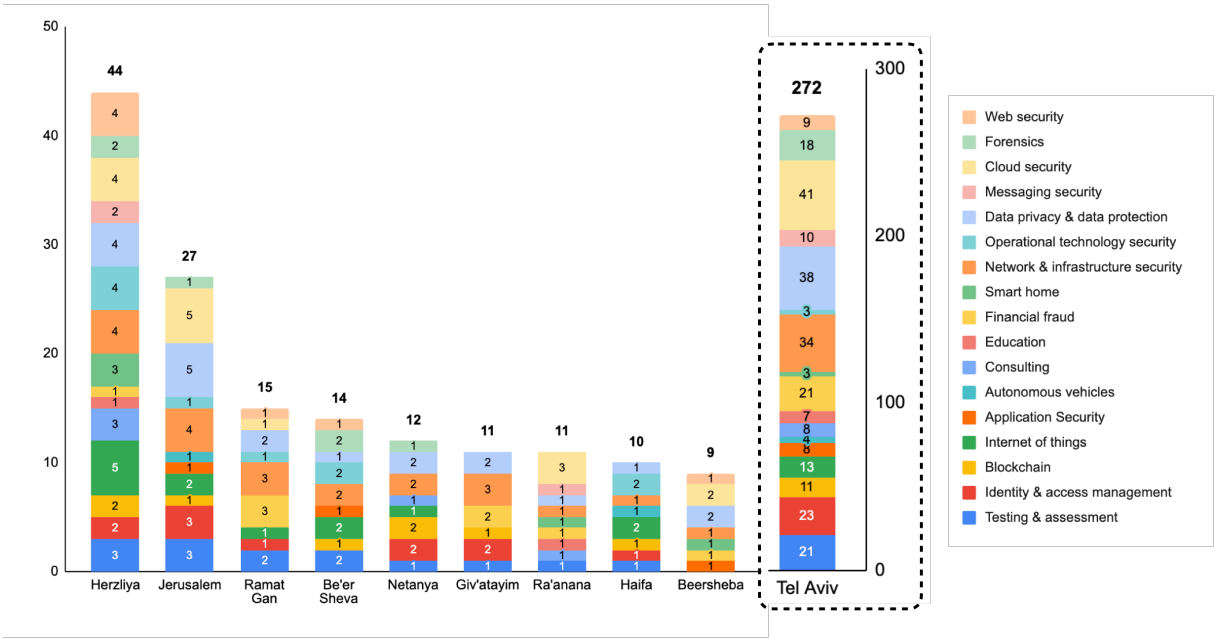
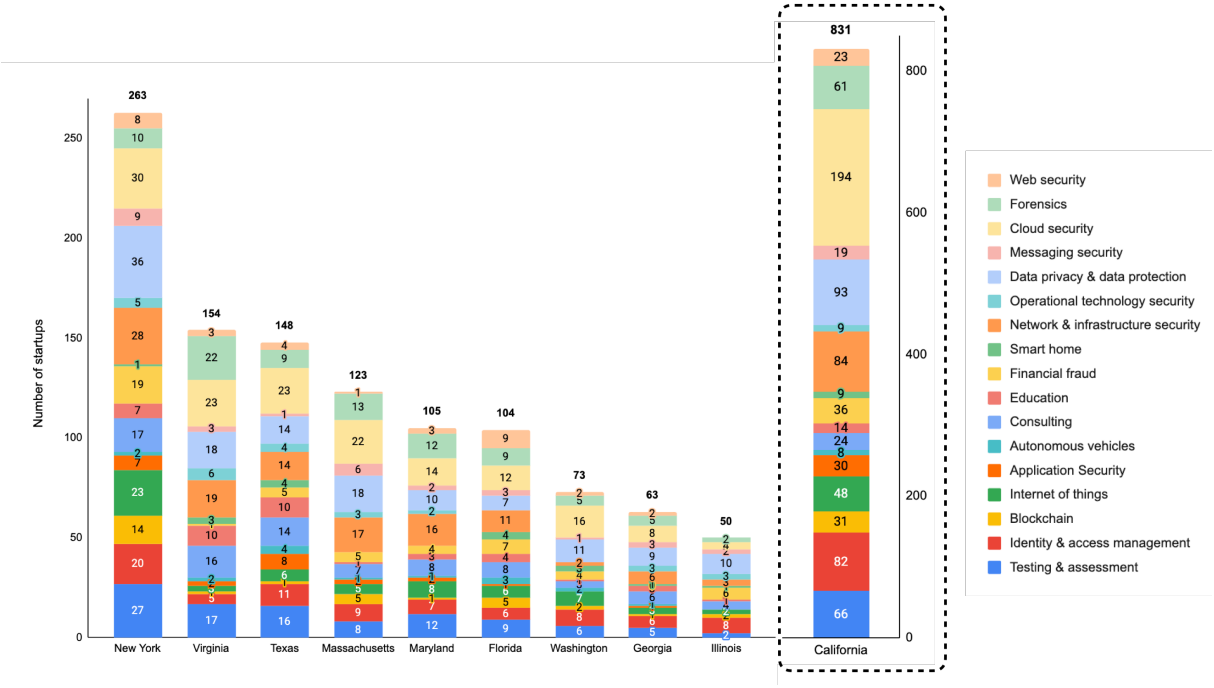


Figure 10

Topic distribution of the Top 10 states in the USA

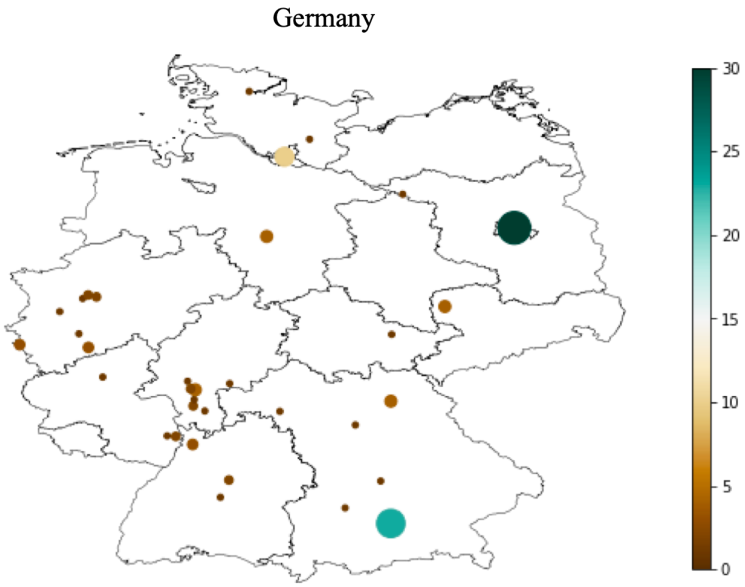


4.4 Entrepreneurial ecosystems

Figure 11 displays the number of startups in every Bundesland in Germany. Bayern and Berlin harbor almost 50% of all cybersecurity startups in Germany. Bremen, Mecklenburg-Vorpommern, Saarland, and Sachsen-Anhalt do not have any cybersecurity startups based on the dataset. The northern and eastern regions around Schleswig-Holstein, Niedersachsen, Sachsen, and Brandenburg also have relatively few startups. Western Germany, including Rheinland-Pfalz, Hessen, Nordrhein-Westfalen, as well as the southern Baden Württemberg, have between six and fifteen startups. The dots in Figure 11 show the cities in which these startups have their headquarters. Many startups are close to the large cities, such as München, Berlin, Cologne, Hamburg, Frankfurt, the metropole region Rhein-Ruhr, and Stuttgart. However, there are also many other cities represented that are not internationally known such as Jena, Regensberg, Polch, and many others. Overall, a relatively clear tendency towards the founding of cybersecurity startups in Germany is observable. The majority of startups are located in the largest cities, Berlin, Hamburg, and Munich, as well as around Frankfurt and the Rhein-Ruhr region encompassing multiple cities.

Figure 11

Location and number of cybersecurity startups in Germany



	Bundesland	Total per state
1	Bayern	31
2	Berlin	30
3	Nordrhein-Westfalen	13
4	Hessen	12
5	Hamburg	10
6	Baden-Württemberg	6
7	Niedersachsen	4
8	Rheinland-Pfalz	4
9	Sachsen	4
10	Schleswig-Holstein	2
11	Brandenburg	1
12	Thüringen	1

Figure 12 provides a worldwide and country-specific (Europe, Israel, USA) overview of cybersecurity startups. It shows that the developed nations have the majority of the businesses. The top 15 out of 93 countries represent roughly 86% of the cybersecurity startups. The USA is an outlier with roughly 1969 startups, followed by Israel (476), the United Kingdom (343), India (193), and Canada (154). The concentration in the other countries is France (103), Germany (102), the Netherlands (100), China (94), Australia (82), Spain (75), Switzerland (64), Singapore (61), Japan (61) and Brazil (52). Developing nations and emerging markets in South East Asia, Africa, and Latin America are occasionally present with less than 5 startups.

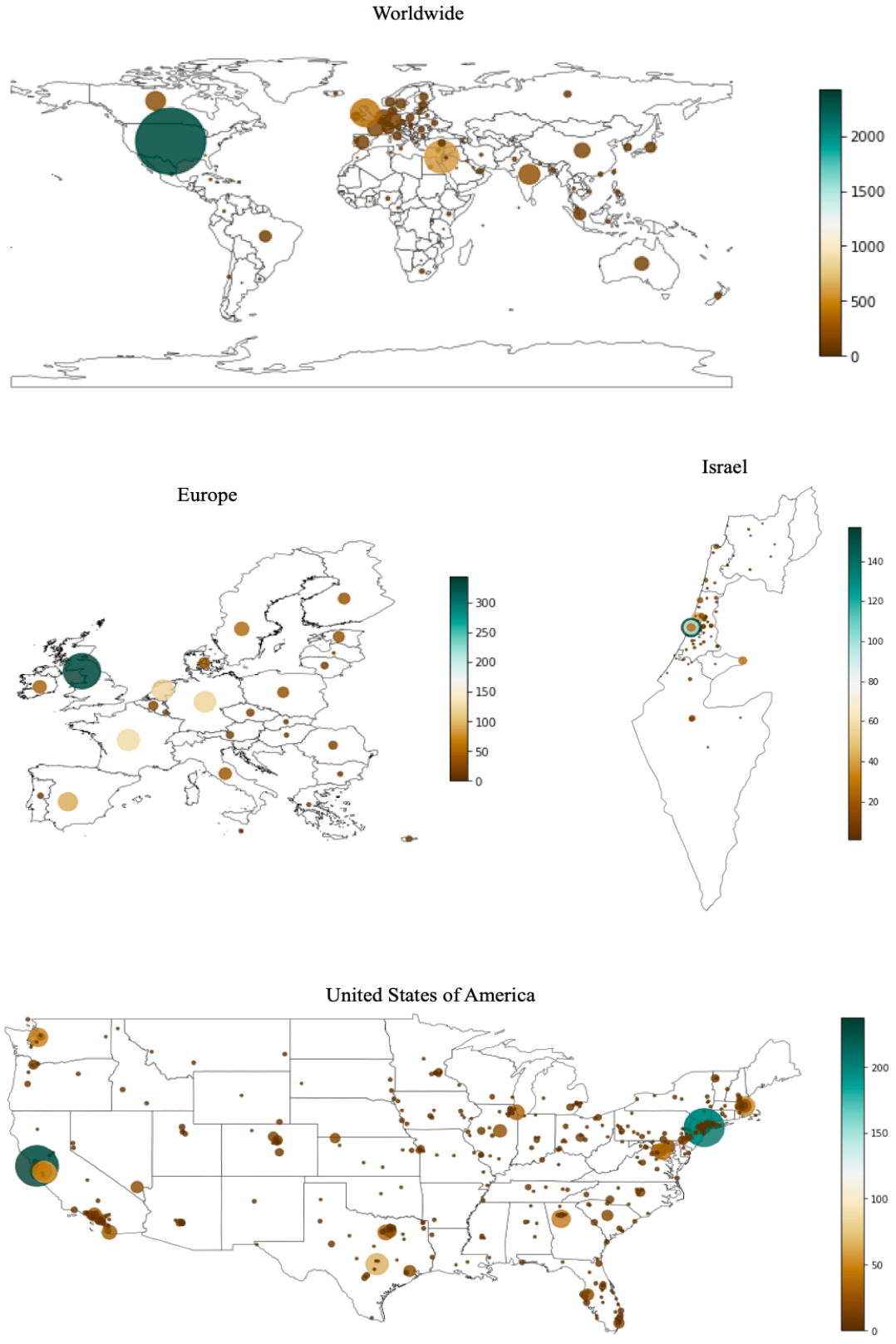
I subdivided Europe's distribution of startups per country into roughly four bins. The first bin constitutes only the United Kingdom because it outputs by far the most cybersecurity startups with more than 330. France, Germany, the Netherlands, and Spain are part of the second bin with a startup range from 75 to 105. The third bin contains countries from the whole of Europe. Especially apparent are the countries in the Nordics, such as Sweden, Denmark, Finland, and Estonia, as well as Poland, Italy, and Ireland. Their number of startups range from 30 to 50. The rest of the European countries have on average below 21 startups within their country, for example, Latvia (2) and Belgium (21)

Israel's cybersecurity startups are mainly concentrated around Tel Aviv, whereas 27 startups are located in the capital Jerusalem. However, some other startups are scattered in the Northern and Southern part of Israel, constituting only less than 10 startups per location. With more than 350 startups around the central part of Israel, it is definitely the main hub for cybersecurity in the country.

Lastly, Figure 12 shows the cybersecurity startup distribution in the USA. The leading hotspot for these startups is California, with a total of 831. Relatively far behind are New York (263), Virginia (154), Texas (148), Massachusetts (123), Maryland (105), and Florida (104). Large parts of the Midwestern-USA show only a small number of cybersecurity startups, on average less than 20 per state. Within California, the main locations for the cybersecurity startups are in San Francisco (237), San Jose (64), Palo Alto (60), Santa Clara (53), Sunnyvale (47), Mountain View (33), Redwood City (27) constituting the Silicon Valley. Another hotspot is in the South around Los Angeles and San Diego. On the East Coast, New York City, Boston and Washington DC exhibit 209, 51 and 40 startups. Other hotspots are in isolation in Austin (65), Seattle (48), Atlanta (45), and Chicago (32).

Figure 12

Location and number of cybersecurity startups worldwide, Europe, Israel, and the USA



5. Discussion

This section aims to answer the research question: *What are the advantages and disadvantages of topic modeling machine learning algorithms such as LDA, CTM, and CorEx for the analysis of entrepreneurial ecosystems?* Hence, I provide an evaluation of the machine learning algorithms for topic modeling to analyze entrepreneurial ecosystems. Next to that, I discuss the best model outcomes to identify the differences and similarities between Europe, particularly Germany, Israel, and the USA. I based the discussion on my logical reasoning combined with domain experts' interviews and academic papers. I use existing literature to a limited extent in evaluating the algorithms because of the pioneering process introduced in this Master Thesis. The majority of existing research in NLP is within the computer science context, whereas the application of these algorithms in a management science context still needs research.

The discussion contains three parts. First, I present an in-depth analysis of the results. I split this analysis into a discussion of the algorithms and the topic distribution of the startups within the geographic focus areas. Second, I explain the limitations of the methodology as well as the outcomes. Last, I illustrate a future direction to analyze entrepreneurial ecosystems with the introduced methodology.

5.1 Discussion of results

For an in-depth understanding of the usefulness of machine learning techniques to analyze entrepreneurial ecosystems, this section primarily deals with the algorithms (LDA, CTM, and CorEx) and does not discuss the geographic locations' outcomes. First, I compare the three algorithms to identify their benefits and advantages for the ecosystem analysis. Second, I illustrate the new approach, which combined CorEx with Word2Vec. Last, I provide an overall statement of the usefulness of the methodology and some best practice guidelines.

5.1.1 Evaluation of the algorithms

I performed a Latent Dirichlet Allocation on two datasets: the dataset from Crunchbase expanded with startups from Startup Nation Central and the scraped text data from the startup's website. The results from both LDAs, in terms of coherence score and topics, underline the fact that automation of the LDA is hard to achieve. For example, the LDA on the description peaks in the experiment with the lowest number of words in the dictionary, whereas the LDA on the website's peaks in the experiment with the most words in the dictionary. This discrepancy does not help to create a rule of thumb to make an industry-independent approach to finding topics. Moreover, the actual formation of topics based on the set of words (see Tables 5 and 6, Section

4.2.1) underlines the unpredictability of unsupervised learning methods. Both approaches displayed interesting words and combinations, but none are entirely coherent and interpretable. This problem is in line with Bianchi et al. (2020), who state that LDA suffers from the sparsity of words and better copes with a large dataset.

The LDA topics based on the startup website indicated at least some coherence, which leads to the suspicion that more text indeed supports the creation of coherent topics. Further exploration of this direction would be interesting. Thereby, the focus should be on scraping only relevant data from the websites instead of simplifying the process by extracting every paragraph. However, the differences in the website structure and the dynamism of many websites make this approach hard to automate. Furthermore, the legal restriction of scraping a website will also persist. In conclusion, the LDA's simplicity allows for a quick and initial exploration of the dataset to get a bird's-eye view of existing terms. Thus, I recommend spending only a limited time optimizing any hyperparameters or the preprocessing steps because of its minimal implication on better topic creation. A quick preprocessing and some experiments to identify the optimal number of topics should be enough to get an initial understanding. For further analysis, more sophisticated approaches that include contextual representation seem to be more promising and, hence, should demand most of the researcher's time and effort.

In order to include the contextualization of words, I used the CTM. This algorithm provided the opportunity to create topics solely on a contextual format or in combination with a bag-of-words approach which is already known from LDA. Based on the results (see Table 7, Section 4.2.2), the combined and contextual versions have outperformed the LDA on the coherence scores. This already indicates that the algorithm better copes with a sparser dictionary than LDA. The algorithms' superiority becomes even more evident when assessing the actual topics (see Table 8, Section 4.2.2), which are easier to interpret and more insightful than the LDA outcomes. Nevertheless, while not every topic provides meaningful insights, the overall performance is better than LDA.

CTM allows the applicant to define the embedding model. The embedding model in this Master Thesis was the "*bert-base-nli-mean-tokens*" which is in line with the embedding model from the authors paper; however, it would be interesting to compare other embedding models and their influence on the topic creation in the future. Moreover, the selection of the embedding model also partially dictates the time needed to run the algorithm. In the end, it is up to the applicant to balance the time-value paradigm in this approach. Overall, the CTM has advantages over LDA, but also some limitations in its application, which I describe in the following.

An advantage of the Contextualized Topic Modeling approach by Bianchi, Terragni, Hovy, Nozza, and Fersini (2020) is its robustness to missing words. In contrast, LDA is initially trained on a specific vocabulary. However, it is possible to get an inadequate representation if the model is applied to a new set of startup descriptions with different words. On the other hand, CTM includes a contextual layer trained on many gigabytes of text data negating this flaw of LDA (F. Bianchi, personal communication [phone interview], 2020). Of course, this is only a limitation in the case of continuous analysis of an ecosystem. For example, new startups are incorporated and added to Crunchbase in relatively short time frames (days or weeks). These changes could influence the outcome of topics in the mid to long-term. Therefore, the LDA model is only useful for a snapshot of the ecosystem. In contrast, CTM could continuously monitor the state of an ecosystem - assumed the researcher wants to apply a trained model and does not want to retrain a model every time.

A weakness of CTM is the requirement for appropriate preprocessing. The authors mention that different kind of preprocessing resulted in different outcomes making it harder to quickly find the best model (F. Bianchi, personal communication [phone interview], 2020). More specifically, through empirical testing, they advise that a vocabulary of around 2000 would most likely create coherent and meaningful topics. This is potentially a limitation when working with a large dataset combining multiple topic fields and hence, increasing the number of words in the vocabulary to an extent the model is not good in handling. This means that the lower the number of words in CTM vocabulary, the easier the reconstruction is while losing some representability of the dataset (S. Terragni, personal communication [phone interview], 2020). On the other hand, the LDA is more forgiving in the preprocessing of the data accelerating the development of topics at the expense of sophistication due to the missing contextual component.

To compensate for the flaws of LDA and CTM, I tested the CorEx algorithm as well. Needless to say, the semi-supervised approach of CorEx is distinct from the other approaches as it is influenceable. The approach to creating topics and the possibility of setting anchor words differentiates CorEx semi-supervised method from the previously discussed unsupervised methods. The anchor words play an essential role in the CorEx method because they direct the model into a "pre-defined" direction. In order to define the anchor words, it is necessary to understand the dataset beforehand and to have relevant domain knowledge. For example, the topics created by LDA and CTM helped with an initial understanding of latent topics in the dataset. Words such as *detection_response, fraud, network, iot, blockchain* and *authentication* hinted towards specific topics. With these terms', it was possible to dive deeper into the cybersecurity space by conducting industry research. The research should be diverse, including

different sources such as market research reports, academic papers, or interviews with industry experts. All of this provides the necessary knowledge to formulate anchor words, understand the created topics, and assess the topics' final distribution. In contrast to the CTM, which is "out-of-vocabulary" resistant due to its pre-trained nature, CorEx has the disadvantage that thought-of anchor words might not exist in the vocabulary. Besides, dealing with abbreviations such as "multifactor authentication" versus "mfa" complicates the process of anchor word creation. These obstacles increase the number of iterations needed to create suitable anchors. Furthermore, I believe that the CorEx approach is somewhat biased because the topics are not explicitly hidden when a priori knowledge is executed.

Although the final topics were more coherent, meaningful, and insightful than the topics created from the other two algorithms, CorEx is dependent on domain knowledge for its application. Since the acquisition of domain knowledge can be time-consuming and could lead to biases based on personal perceptions of topic direction, I created a more independent approach. Hence, I introduced a combination of CorEx anchor words based on domain knowledge with contextualized embeddings for similar words by Word2Vec. I discuss this new method in the next section.

5.1.2 Semi-supervised learning doped with Word2Vec

The requirement of domain knowledge to create meaningful anchor words in CorEx makes its application not as fast as other approaches. Therefore, I introduced a semi-automated approach coined "CorEx doped with Word2Vec". This approach's unique feature is an enhancement through the inclusion of similar words based on distances of the word embeddings. More specifically, this approach provides the applicant with a tool to define non-obvious anchor words to facilitate the creation of meaningful topics.

One of the difficulties in CorEx was the definition of anchor words that existed in the vocabulary. The Word2Vec model provides a method to access the most similar words to an input. This input could be a single word such as *authentication* or a set of words such as *[authentication, password, biometric]*. In general, Word2Vec turns text into a numerical representation called vectors and puts them into a vector space. "With enough data, usage and contexts, Word2Vec makes accurate guesses about a word's meaning based on past appearances" and suggests similar words based on the cosine similarity (Nicholson, n.a., p. 1). With this approach, Word2Vec makes it possible to find meaningful anchor words that support CorEx in creating insightful topics. Moreover, the combination with Word2Vec provides CorEx with a contextual element that it does not have inherently. Similarly, Moody (2016) introduced

a combination of Word2Vec with LDA. The author explains that the approach "allows for unsupervised document representations [...] while simultaneously learning word vectors and the linear relationships between them" (Moody, 2016, p. 1). It combines the largely uninterpretable word vectors from Word2Vec with the interpretable LDA that misses local world relationships. Although the author describes the lda2vec method as a mostly experimental approach, it is possible to draw some parallels to the CorEx infused Word2Vec approach introduced in this Thesis. Of course, these two approaches are different as one provides an integrated framework with source code (lda2vec), whereas the other arguments the application side through separate enhancement (this Thesis). Nevertheless, the introduction of the "CorEx doped with Word2Vec" supported the creation of coherent and meaningful topics by providing a clearer picture of words in the vocabulary and decreasing the time needed for research. After evaluating the algorithm, the next section dives deeper into more general best practices when applying the approach of this Master Thesis.

5.1.3 Guideline for an industry-independent approach to analyze entrepreneurial ecosystems

The usage of unstructured data, such as text, is still mainly at its infant steps. However, research in Natural Language Processing has accelerated over the years ranging from automatic summarization and question answering to text-to-speech and topic modeling. In this Master Thesis, I borrowed algorithms from the NLP subfield of topic modeling to introduce the techniques in a management research context. Based on this Thesis outcomes, it becomes apparent that this data-driven methodology provides valuable insights to analyze entrepreneurial ecosystems independent of a specific industry. In the following, I define some best practices upon which other researchers might rely to accelerate and simplify the application of topic modeling. Figure 13 also visualizes these guidelines for a better understanding of the flow.

The first part, and probably the most crucial part, is the dataset. Quality and quantity are the two dimensions that define a good dataset in machine learning applications (Al-Jarrah, Yoo, Muhaidat, Karagiannidis, & Taha, 2015). Crunchbase provides a good solution to create a dataset because of its large database. You could also extend the dataset with other databases to increase the number of companies; in this Thesis, I used the database of Startup Nation Central. However, you should always check the dataset for duplicates and other measures. For example, a histogram of the descriptions' length helps to exclude companies with a short description right from the beginning.

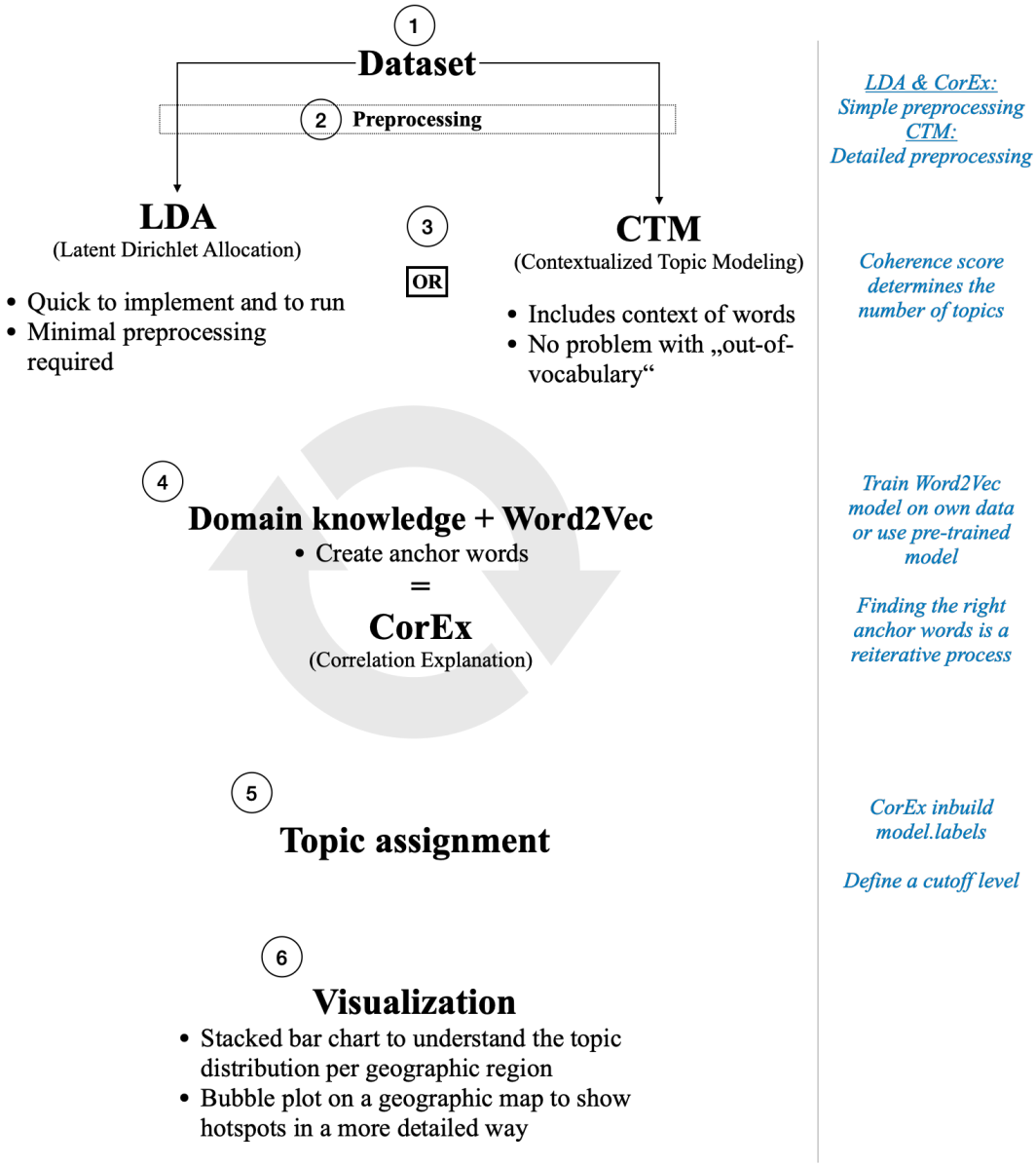
The second part includes an adequate preprocessing to prepare the data for further analysis. In the case of LDA and CorEx a simple preprocessing seems to be enough, whereas CTM requires more sophisticated preparation to provide an adequate analysis. The third step is about a quick execution of the unsupervised methods to get a rough birds-eye view of the topic direction. For that, it is essential to run the algorithms multiple times to identify the optimal number of topics based on the coherence score. According to the experiments run in this Thesis, LDA is easy to implement and to run. In contrast, CTM is more sophisticated because it includes the context of words and deals with the "out-of-vocabulary" problem. It is up to the applicant to decide which of these two approaches to use in the beginning. From this stage, further research to gain some domain knowledge is beneficial.

In the fourth step, you should combine the domain knowledge with the word embeddings from Word2Vec to create anchor words for the CorEx model. This is the most time-consuming step because of the reiterative nature of creating anchor words that facilitate meaningful topics. Moreover, the applicant must decide if it makes sense to train a Word2Vec model on their own dataset or use a pre-trained model. In the fifth step, you must assign the topic to their respective startups. CorEx provides an in-build "model.labels" method that automatically assigns single, multiple or no topics to a startup. This method sometimes works to a limited extent, meaning that a large chunk of companies does not get a topic allocated. In this case, you should define a threshold for the mutual information score for the topic assignment manually by drawing some samples and determining a cutoff level.

After the assignment, various possibilities exist to visualize the results for a better understanding. For instance, stacked bar charts provide effective means to visualize the topic distribution per region comparatively. Besides, a bubble plot on a map shows the hotspots more detailed than a normal choropleth map. After this step, it is possible to draw inferences about the state of the cybersecurity ecosystem within the geographical areas. Thus, the next section aims to answer the second research question: *Resulting from the best topic modeling approach evaluated in Research Question 1, what are the differences and similarities between Europe (in particular Germany), Israel, and the USA in terms of their cybersecurity ecosystem?*

Figure 13

Best practice approach to create topics



5.2 The entrepreneurial ecosystem of the cybersecurity industry

The previous section focused on evaluating topic modeling algorithms to analyze entrepreneurial ecosystems, primarily focusing on the methodological approach to create an industry-independent approach. This section discusses the similarities and differences in the entrepreneurial hotspots of the cybersecurity industry based on this approach. Hence, the discussion revolves around Europe, Israel, and the USA, as well as an isolated observation of Germany.

5.2.1 Comparison of the entrepreneurial ecosystem in Europe, Israel and the USA

Based on the assigned topics to the startups, it was possible to analyze the distribution and focus of startups in Europe, Israel, and the USA. Section 4.3 and 4.4 already displayed the relative and absolute distribution of topics as well as illustrated entrepreneurial hotspots in these areas. In the following, I discuss similarities and differences to explain the findings.

The occurrence of the majority of startups in Europe, Israel, and the USA, as shown in the worldwide view in Figure 12, does not come as a surprise. This is because these locations are also spearheading the worldwide total venture capital investments (Statista estimates, 2019). According to (Press, 2017), Israel is widely known to produce an array of cybersecurity companies for several reasons. For example, Israel fosters collaboration between the government, businesses, and universities as it views the industry as an economic growth engine. Moreover, the government tries to play a guiding role while attempting to remove too much interference to reduce the tension seen when governments work hand in hand with cybersecurity companies. For that, Israel quickly iterated through multiple operational structures by introducing establishments such as the National Cyber Bureau in 2011 or the National Cyber Security Authority in 2015. Additionally, the geopolitical situation of Israel puts its innovation focus on the military sector. Over the years, the Israeli Defense Unit 8200 evolved into an innovation powerhouse focusing on cybersecurity and other fields. Young people that serve in the military, for example, in the 8200 unit, gather real-life experience and work on leading-edge cybersecurity challenges and solutions throughout this phase (Raska, 2015). Afterward, they are well-equipped to start their own venture. Such a path is entirely missing in Europe and the USA.

The nation has formed numerous partnerships that outline the importance of Israel's cybersecurity initiatives. For example, the Israel-Europe R&D Directorate tries to increase the scientific and industrial collaboration between Israel and the European research and innovation ecosystem. Such an initiative supports Israeli companies in partnering with foreign companies on projects and ultimately extending their reach (Aharon, 2018). Lastly, top technology companies like Microsoft, Google, Intel, or Oracle that perform R&D in Israel fuel Israel's innovation ecosystem (Minevich, 2020). According to many researchers, this tight-knit ecosystem of innovators, government, domestic and foreign companies, and universities is the key differentiator of Israel's success in cybersecurity and other industries (Adamsky, 2017; Tabansky & Ben Israel, 2015).

However, although Israel founds many innovative companies, they often move to the USA relatively fast. The USA makes up almost 50% of the total cybersecurity market worldwide and

cannot be ignored when creating a tech company (Khalid, 2017). Although it is not a complete move to another country because Israel mostly keeps R&D inland, it conflates the US and Israel ecosystem. Due to the importance of cybersecurity for protecting critical infrastructure and cyberattacks from hostile groups or other nations, the US Department of Homeland Security has published a cybersecurity strategy in 2018 (U.S. Department of Homeland Security, 2018). In contrast to Israel, the US government does neither seem to take an active investor role nor actively support cybersecurity innovation. The relationship between startups and the government appears to be more distant than collaborative. However, the country's robust venture capital ecosystem acts as an innovation financer next to larger corporations.

Europe's cybersecurity ecosystem is very fragmented, with France, the Netherlands, and Germany as the states with the highest output of startups (see Figure 8). Research shows that European companies rarely look for cross-border international partnerships or customers (Vaugan, 2020). The process often takes place much later when the risk of failure diminishes. Unfortunately, this is not only the case for the cybersecurity industry but the European startup ecosystem in general. This is different for Israeli companies with such a small domestic market that they must go international as quickly as possible. US companies are already in the largest market; hence they do not have to consider the internationalization process in the beginning. Although cybersecurity is often a priority for national governments, the EU plans for a pan European cybersecurity center to share knowledge, competences, and capital to foster innovation and create successful cybersecurity companies (Naujokaitytė, 2020). It remains open how such initiatives will influence the output of cybersecurity companies and their international acknowledgment and success.

5.2.2 Topic distribution in Europe, Israel, and the USA

Section 4.3 revealed that a trend towards specific topics becomes apparent in all four geographic arrangements. These topics are *Data privacy & data protection* in Germany; *Network & infrastructure security* and *Data privacy & data protection* in the European Union; *Cloud security, Data privacy & data protection* and *Network & infrastructure security* in Israel and the USA. In this section, the Master Thesis tries to explain the differences and similarities to create a context.

The movement to cloud computing serves as a catalysator for higher efficiency and productivity in companies while at the same time accelerated the need for cybersecurity solutions (Jathanna & Jagli, 2017). In its essence, third-party data centers "station" the cloud and make it available to a multitude of users online 24/7. The increasing reliance on cloud computing through third-

party solutions has also introduced the question of data privacy & data protection (Ismail, 2018). A response to that question was the introduction of the EU General Data Protection Regulation (GDPR) and a not as strict version in California, the California Consumer Privacy Act (CCPA). This interaction between the introduction of cloud computing and government regulation to protect consumers gave rise to more innovation in the data privacy & data protection space, also indicated by the number of startups in this topic (see Figures 7, 8, 9, and 10).

Gartner (2020a) expects the cloud service market to increase by 18,4% to almost $ 305bn in 2021. This steady increase underlines the growing market for cloud adoption by businesses and endorses the necessity for cloud security solutions reflected in the high number of startups in Europe, Israel, and the USA that focus on this topic. In contrast to the high number of startups dealing with cloud security is the relatively small amount invested into this segment ($ 585m in 2020) compared to other segments such as infrastructure security ($ 17,4bn) or data security ($ 2,8bn) (Gartner, 2020b). Remarkably, the growth rate from the previous year in cloud security is more than 33,3%, compared to merely 5,8% and 7,2% for infrastructure and data security, respectively (Gartner, 2020b). This difference needs further investigation by drawing samples from the cloud security startups to confirm their inherent focus on cloud security rather than merely offering another security solution that works in the cloud. Nevertheless, since the emerging of cloud solutions started in the US, it makes sense that the US also leads the cloud adoption worldwide. Gartner (2019) illustrates the adoption rate in terms of total IT spending on cloud services. On this chart, Germany classifies as a lagging country next to other European countries such as France, Italy or Spain. The slower adoption rate of cloud solutions might explain the lower number of cloud security solutions in Europe compared to the US. However, dedicated research on the correlation of cloud adoption and the founding of cloud security companies does not exist.

Germany, Europe, Israel, and the USA strongly represent the topic of *Network & infrastructure security*. With its large market size, it makes sense that multiple startups are active in this space. However, *Operational technology security,* which might overlap with *Network & infrastructure security* when inspecting critical infrastructure more closely, is not so present in the countries. Operational technology describes the systems used to run manufacturing plants, control power stations, and water utilities which often describes critical national infrastructure. According to Mansfield-Devine (2019), these systems are often poorly protected, although cyber-attacks are persistent and often detrimental. Hence, it is surprising that only a small number of startups can be associated with the *Operational technology security* topic in the dataset. However, some

companies dealing with critical infrastructure security and operational technology may be included in the *Network & infrastructure security* topic as well.

The topic *Forensics* often rereferred to as digital forensics, deals with the aftermath of a cyber-attack. The growing threat landscape and increase in cyber-attacks also accelerate the market growth of digital forensics solutions and services (Mordor Intelligence LLP, 2020). Bieringer (personal communication [phone interview], 2020), the Head of Entrepreneurship & Technology Transfer at the CISPA, representing an institute for cybersecurity in Germany, identifies forensics solutions as an essential pillar in cyberspace going forward. Based on the results in Table 10 (Section 4.3), it becomes evident that Germany is lagging behind Israel and the USA in this trending field of cybersecurity. This insight is interesting for cybersecurity professionals that are interested in founding a startup and venture capital investors. It provides professionals with the opportunity to develop digital forensics solutions and investors with the possibility to finance an innovative and upcoming segment.

Europe, Germany, Israel, and the USA do not significantly differ in topics such as *Autonomous vehicles*, *Web security*, and *Messaging security*. Moreover, these topics make up only a small share of the overall number of startups in these regions. An explanation for the low number is the nature of these segments. The development of autonomous vehicles is still mostly behind the hype that came with it because of technological reasons and government regulation (Fagella, 2020). Therefore, it makes sense that the security efforts in this field cannot be as extensive. Moreover, the success of autonomous vehicles depends on their safety, both from a manufacturing point of view but also from a cybersecurity point. Hence, it makes sense that the car manufacturer provides its own safety systems rather than relying on third-party solutions from startups. In addition to that, you can observe the topic of web security from a different angle. Since the introduction of the world wide web in the 90s, web security has played an important role for a very long time. Therefore, it makes sense to assume that already incumbent players are responsible for the majority of web security. Thus, the field is not so attractive for startups anymore, explaining the relatively low attribution of web security startups in the dataset.

As an application's underlying technology, *Blockchain* is not as widely distributed as other topics except in California, USA. The overall low attribution towards Blockchain technology probably results from its relative newness and the low number of real use cases. It was only four years ago (2016) that the World Economic Forum coined Blockchain as one of the top 10 emerging technologies (World Economic Forum, 2016). Already before that, fans glorified Blockchain as a technology that will disrupt every industry. However, at the current state of the

technology, the market distinguishes between theoretical and practical use cases and already implemented solutions are still having limitations (Friedlmaier, Tumasjan, & Welpe, 2018). Hence, it remains open how practitioners adopt the technology in the cybersecurity space, especially when the technology in itself is supposed to have an inherent safety feature with its immutable decentralized ledger principle. Still, a study by Friedlmaier et al. (2018) is in accordance with the outcomes from the results in this Master Thesis, that the USA has the highest density of blockchain startups. Their study goes a step further and indicates that the US covers the total funding of blockchain startups by 50%.

Overall, this section shows that further in-depth analysis is possible with the assignment of topics through machine learning algorithms. The creation of topics and their assignment to each company in the dataset provides the possibility to compare topics on a regional level. It is possible to identify pioneering and lagging countries. Moreover, it gives stakeholders thorough insights into the current state of an industry. For instance, entrepreneurs get an overview of international competition and venture capital investors can identify potentially underfunded or oversubscribed segments.

5.2.3 The potential of Germany as an entrepreneurial hotspot in cybersecurity

I split this section into two parts to identify the current state and the potential of Germany as an entrepreneurial hotspot in cybersecurity. The first part discusses alignments and discrepancies of the key areas in cybersecurity identified by the Federal Ministry of Education and Research as well as a comparison of the topic distribution between the Bundesländer. The second part assesses the potential of Germany as an entrepreneurial hotspot for cybersecurity with a SWOT analysis. By identifying strengths, weaknesses, opportunities, and threats, it is possible to determine the potential for improvement.

5.2.3.1 Current state of the German cybersecurity ecosystem

The Federal Ministry of Education and Research identified four key cybersecurity areas, namely industry 4.0, privacy, critical infrastructure, and cloud computing, to secure and enhance Germany's position (Bundesministerium für Bildung und Forschung (BMBF), n.d.-a). Based on the analysis in section 4.3, it is possible to confirm that most startups focus on these areas. More specifically, 15 companies cover *Data privacy & data protection* and 13 companies cover *Cloud security*. Industry 4.0 and critical infrastructure are not so easy to pinpoint because they fall mainly in *Network & infrastructure security* (13) and *Operational technology security* (4), which are not solely specific to protect critical infrastructure but network structures and

physical devices in general. Nevertheless, the analysis identifies an opportunity in the space of Operational technology because of the increasing connectedness of machines and critical infrastructure to the internet.

The distribution of the topics from a Bundesland perspective does not provide any meaningful insights because of the low numbers (see Figure 7, Section 4.3). Bayern and Berlin are the locations with the greatest number of startups; however, this is probably associated with reasons that apply to all startup industries, such as proximity to investors, a pool of international talent, and the attractiveness of the city. However, an infographic shown in a FactSheet of the Germany Trade & Invest (GTAI) maps IT Security Hubs and R&D institutes mostly in Hessen and Nordrhein-Westfalen (Germany Trade and Invest, 2019). There seems to be a discrepancy between fundamental research and the application of that research in these regions. In future research, it would be interesting to inspect the type of companies founded in these regions even closer to potentially differentiate between ground-breaking innovation coming from these institutes and more applied solutions in Berlin or Bayern.

The only noticeable difference in Figure 7 is the high number of *Data privacy & data protection* startups in Berlin (7) compared to Bayern (3) and the other Bundesländer. Other than that, it is not possible to identify topic-based hotspots in Germany. A more in-depth analysis, for example, on the distribution of B2C, B2B, or B2G companies is also not possible with the data points available from the dataset. However, due to the nature of cybersecurity and the importance for businesses, I expect that the majority of companies focus on B2B.

5.2.3.2 SWOT analysis of the German cybersecurity ecosystem

A SWOT analysis helps assess the current position of an industry and supports the decision-making process of a new direction (Helms & Nixon, 2010). With the SWOT analysis of the German cybersecurity industry, it is possible to put the previous findings into context and to evaluate the potential of that ecosystem. Therefore, I conducted a SWOT analysis of the German cybersecurity industry in the following.

Strengths:

Since 2011, the BMBF has been supporting three competence centers with a focus on IT security. These are CISPA in Saarbrücken, EC_SPRIDE in Darmstadt, and KASTEL in Karlsruhe as well as the Ruhr-Universität in Bochum. They have a consulting and support function to help interested founders to develop and evaluate their idea up to market entry (Bundesministerium für Bildung und Forschung (BMBF), n.d.-b). Such initiatives signal strong

support from the government towards the development of a cybersecurity ecosystem in Germany. A positive aspect is that the government and the institutes create the setting for an ecosystem's organic development rather than a substantial interference.
Moreover, Germany has a network of digital hubs far more distributed than France and the UK, which solely have startups consolidated in Paris and London. As already shown in Figure 11 (Section 4.3), the high quantity cities Berlin and Munich are complemented by regional hubs in Hessen, Nordrhein-Westfalen and Hamburg. The competence centers, universities, SMEs and corporations support these locations by serving as pilot customers and partners. Furthermore, lower rents compared to the major cities and access to specialized talent from local universities as well as international talent support the specialization of these hubs (L. Bieringer, personal communication [phone interview], 2020). Thus, I expect more and more successful startups to spring from the regional hubs rather than the existing hotspots of Munich and Berlin.

Weaknesses:
At the moment, the translation of academic research into a real-life application through spin-offs is relatively low. This is often due to wrong incentive structures within universities or institutes such as the Fraunhofer Institute as well as economic and bureaucracy hurdles. According to a newspaper article, the Fraunhofer institute was widely criticized for taking a high equity share, forcing licensing fees and revenue share from spin-offs (Stölzel, 2020). This setting discourages spin-offs and drives away serious investors to further support ideas. It is necessary to align the interest of research institutes with the need of the market.
The broad surface of cybersecurity attacks poses an overall challenge for businesses. Many attacks happen secretly, and victims often do not see the impact of the attacks or the IT security team's efforts. This translates into an awareness problem within companies. According to a study by Dreißigsacker, von Skarczinski, and Wollinger (2020), business executives estimate the likelihood of a random and targeted cyber-attack to be lower than their IT security team. Other employees estimate the probability of an attack even significantly lesser than the IT team and executives. However, such a discrepancy is a weakness within a company because employees can be targeted with phishing emails or the CEO-Fraud technique. Additionally, business executives are responsible for allocating budget. Therefore, lower awareness of the risks might translate into lower IT security spending. Thus, currently, the factor cybersecurity awareness is a problem within businesses. This weakness differentiates between companies of different sizes, but overall it is necessary to educate employees continually and business executives to take cybersecurity seriously and do not judge the risks as unlikely to happen.

Opportunities:

There are many opportunities for the German cybersecurity landscape. According to a study by Germany Trade and Invest (2019), SMEs are only at the start of their digital transformation and turning increasingly to cloud solutions. Additionally, they perceive IT security and compliance issues as the most significant barriers to a successful integration, which provides opportunities for cybersecurity startups. With the European General Data Protection Regulation introduction in 2018 (European Commission, 2018), the German cybersecurity market should directly benefit from increased demand from domestic and international companies. The EU forces domestic companies to follow the regulation and requires international companies that want to do business in the EU to follow suit. In total, the global cybersecurity market has grown over the last years and is expected to grow over the next years. Research puts the sector's compound annual growth rate at 10% from 2020 to 2027 (Grand View Research, 2020). This growth provides the opportunity for German startups to scale in Europe and internationally.

A more general opportunity is the fast pace in cybersecurity, which screams for constant innovation. Constant development in the software and hardware space often comes at the expense of sophisticated security. This phenomenon is observable worldwide, which creates the chance to close the gap with innovative solutions.

Threats:

The dependency of German startups on foreign investors, especially in later funding rounds or asset-heavy models, might hurt Germany in the long-term in this globalized world. It does not matter where the money comes from for the startup, but it might threaten the international competency on an ecosystem and country-level of the region (Wijngaarde, 2020). Moreover, German corporates (DAX 30) which have the resources to invest in domestic startups, only spend a fraction of their innovation budget on external innovation. A report by Hilpert, Meermann, and von Borries (2019) shows that DAX 30 corporates invest around 3,3% of their revenue in innovation, from which 96% falls on internal innovation and only 4% (€ 16bn) on external innovation. This stands in stark contrast to its counterpart in the USA and China, which invest 2x and 12x more compared to the DAX 30. These low numbers are disturbing when considering that cybersecurity is only a single industry upon which these investments fall. The willingness to invest and to acquire domestic cybersecurity startups is essential to develop a sustainable ecosystem. Hence, the low adoption and acquisition rate threatens the development of a thriving cybersecurity ecosystem.

Moreover, the already high number of large cybersecurity firms from overseas and international investors, that are better funded and that engage in growth-stage funding, strengthen this shortcoming. Hence, larger overseas players that have already built trust and entered the German market have an advantage over domestic startups. Additionally, big tech firms such as Google, Microsoft, Amazon, or Apple already bundle cybersecurity as part of their offerings, making it harder for German startups to tag along as third-party providers (Pierre Audoin Consultants, 2013). Overall, stricter regulation on software and hardware companies in terms of cybersecurity could lead to a shift from external solutions (e.g. from startups) to more sophisticated in-build solutions at the product's source.

5.3 Limitations

Researchers have not researched and applied the field of Natural Language Processing in a management context widely. Hence, it is natural that limitations to the method exist as the work is at its infant stage. Throughout the Master Thesis, some limitations became apparent which I explain in the following.

First, the analysis of the cybersecurity entrepreneurial ecosystem is highly dependent on the representativeness of the data. Hence, the primary usage of a single database (Crunchbase) potentially distorts the overall analysis to some extent. Not every company defined under productive entrepreneurship is accessible via Crunchbase. Especially younger companies whose future is unclear are most likely not to be present within the database. This limits the approach to a status-quo analysis instead of a tool to identify hidden or emerging trends. Thus, the combination of multiple databases and the scraping of directories or websites to increase the number of companies in the dataset is a potential direction to evaluate in the future.

Second, the nature of unsupervised machine learning algorithms such as LDA and CTM creates difficulties in accurately generating and assessing the topics. The topics' assessment is not automated; hence, some room for human error exists. Moreover, the assessment step is highly subjective and depends on a certain degree on available domain knowledge. Additionally, there is no objective score to measure the accuracy of the total number of topics nor the correctness of assignments on a single company level. Overall, the variability of hyperparameters, the preprocessing steps, and the limited evaluation options make an automatic application of the process still unlikely.

Third, the comparison of Europe, in particular Germany, to Israel and the USA is mostly based on a quantitative approach in terms of the absolute and relative focus of the startups within these regions. This might create a view that the locations with the highest number of startups are also

at the forefront of cybersecurity. However, the number of startups is not a measure for a specific topic's quality and future growth. Thus, the reader must carefully interpret the explanatory power of the comparison due to the unequal distribution of startups in the regions. Future work should include other parameters, such as total funding amount, number of employees, revenue, and website visits if the necessary data is accessible.

5.4 Future direction

In this Master Thesis, the focus was on unsupervised and semi-supervised learning. The main reason for that was the nature of the unlabeled data. It would be interesting to either acquire (if existent) or manually label a certain number of data points (startup descriptions) with a specific topic to train a model that utilizes supervised learning. Since supervised learning provides much more accurate results than unsupervised learning verified by the accuracy score, this approach potentially assigns more accurate topics to the startups resulting in a more precise analysis. However, a hurdle to overcome is the high time-investment in manually labeling the data or a high monetary cost in purchasing labeled data.

Furthermore, I combined the CorEx approach with similar words from Word2Vec. Word2Vec is an unsupervised approach based on a distributional hypothesis. This means that words that occur in the same context tend to have a similar meaning (Weaver, 1955). The Word2Vec embeddings have seen further development to provide even better outcomes. Most notable are FastText (extension of Word2Vec) and ELMo. FastText improves the Word2Vec approach by allowing the computation of words that did not appear in the training data, solving the limitation of "out-of-vocabulary" words (Bojanowski, Grave, Joulin, & Mikolov, 2017). For future research, it would be interesting to compare the outcomes of other methods, such as FastText and ELMo, to define even better anchor words than with Word2Vec.

Last, I already mentioned the primary reliance on one data source (Crunchbase) in the limitation section. In the future, it would be interesting to combine even more databases to generate a complete picture of the industry in focus. This requires the existence of and access to such databases. Moreover, accessing different databases most likely poses the challenge of having different data types that need work to become homogeneous for further processing. Suppose the possibility exists to access data of recently founded startups, for example, through registry or university scraping. In that case, a more future-oriented outlook could extend the analysis of the status quo of the entrepreneurial ecosystem.

6. Conclusion

This Master Thesis aimed to evaluate machine learning algorithms for the analysis of entrepreneurial ecosystems. Based on that evaluation and the cybersecurity industry as a showcase, I developed recommendations for applying the algorithms for an industry-independent approach. Existing research focused primarily on identifying and measuring the entrepreneurial ecosystem rather than an in-depth analysis of the regional state of a specific industry. I tried to close this gap with this Master Thesis. With the utilization of machine learning algorithms, I introduced a new approach that is more data-driven, resulting in a faster and more automated analysis.

The outcomes of the topic modeling algorithms Latent Dirichlet Allocation, Contextualized Topic Modeling and Correlation Explanation illustrate the viability of the approach in analyzing entrepreneurial ecosystems. Although substantial differences in coherence and insightfulness of the topics based on the unsupervised and semi-supervised methods exist, they provide the basis for an initial understanding (through LDA and CTM) as well as a more detailed understanding of an entrepreneurial ecosystem (through CorEx).

In this Thesis, it became apparent that the usage of the algorithms depends on the required sophistication of the outcome. For example, the relatively fast and easy implementation of LDA, which requires minimal preprocessing at the expense of context interpretation, enables an initial understanding of topic direction. The creation of coherent topics with this approach turns out to be unlikely, yet it fabricates a basis for further exploration with more sophisticated algorithms. The CTM counteracts the biggest flaw of LDA and CorEx by including the context of the words in creating topics and models that apply to newly added startup descriptions due to its "out-of-vocabulary" resistance. Through word embeddings, the algorithm can create more coherent topics than LDA. Nevertheless, these topics are still not as coherent as required for a suitable assignment to companies and a later analysis. CorEx triumphs the other two approaches by facilitating a semi-supervised learning approach that enables the researcher's input and interference. The setting of anchor words guides the model in a pre-defined direction, enabling the creation of typically underrepresented topics that were most likely not uncovered by LDA or CTM. However, the required domain knowledge as well as the missing context component in the formulation of topics is a downside in CorEx. Thus, I introduced a new approach in this Master Thesis to diminish these shortcomings.

The new approach combines the strength of manually setting anchors words based on domain knowledge with the contextualized word embeddings from Word2Vec. Word2Vec provides the possibility to create a model based on own data or the use of a pre-trained model. The model

enables access to the most similar words based on an input, such as a word associated with a topic from the domain knowledge. This addition makes the creation of the topics by CorEx more robust because the Word2Vec gives access to abbreviations and related words that could otherwise be missed. Thus, CorEx combined with Word2Vec provided the most coherent and insightful topics which I then assigned to the respective startups in the dataset. Based on the assignment, further analysis of the topic distribution and entrepreneurial hotspots in Europe with a particular focus on Germany, Israel, and the US was possible.

The analysis of the cybersecurity industry provided a showcase for the methodology. In general, the introduced approach is suitable for every other industry. The discussion of the results shows that it is possible to gain an overview of the startup landscape on a regional basis. From that standpoint, it is feasible to dive deeper into differences and similarities with a qualitative analysis. For example, the results indicated the high number of startups with the topic of *Data privacy & data protection*, *Network & infrastructure security*, and *Cloud security* as well as underrepresented topics that are on the rise, such as *Digital forensics*. With further research, it was possible to attribute reasons and justifications to their occurrence and appearance.

In conclusion, this Master Thesis recommends researchers and practitioners to include the introduced machine learning approach to extend their current entrepreneurial ecosystem research methods. While scientists research the machine learning algorithms in Natural Language Processing extensively in a computer and data science context, this should not keep researchers from their application in other domains such as a management science context. With the constant innovation in NLP, the prospects for more in-depth and data-driven analysis of entrepreneurial ecosystems look more than promising.

References

Accenture. (2020). *Innovate for cyber resilience*. Retrieved from https://www.accenture.com/_acnmedia/PDF-116/Accenture-Cybersecurity-Report-2020.pdf

Ács, Z. J., Autio, E., & Szerb, L. (2014). National Systems of Entrepreneurship: Measurement issues and policy implications. *Research Policy, 43*(3), 476-494. doi:10.1016/j.respol.2013.08.016

Adamsky, D. (2017). The Israeli Odyssey toward Its National Cyber Security Strategy. *The Washington Quarterly, 40*(2), 113-127.

Aharon, A. (2018). How Israel Is Accelerating Cybersecurity Innovation. Retrieved from https://blogs.timesofisrael.com/how-israel-is-accelerating-cybersecurity-innovation/

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research, 2*(3), 87-93.

Amornsiripanitch, N., Gompers, P. A., & Xuan, Y. (2019). More than money: Venture capitalists on boards. *The Journal of Law, Economics, and Organization, 35*(3), 513-543.

Audretsch, D. B. (2014). From the entrepreneurial university to the university for the entrepreneurial society. *Journal of Technology Transfer, 39*(3), 313-321. doi:10.1007/s10961-012-9288-1

Audretsch, D. B., Keilbach, M. C., & Lehmann, E. E. (2006). *Entrepreneurship and economic growth*: Oxford University Press.

Audretsch, D. B., Lehmann, E. E., & Warning, S. (2017). University spillovers and new firm location. In *Universities and the Entrepreneurial Ecosystem*: Edward Elgar Publishing.

Auerswald, P. E. (2015). Enabling entrepreneurial ecosystems: Insights from ecology to inform effective entrepreneurship policy. *Kauffman Foundation Research Series on city, metro, and regional entrepreneurship*.

Baumol, W. J. (1994). *Entrepreneurship, Management and the Structure of Payoff* MIT Press.

Becattini, G. (2004). *Industrial districts: A new approach to industrial change*: Edward Elgar Publishing.

Bianchi, F., Terragni, S., & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2020). Cross-lingual Contextualized Topic Models with Zero-shot Learning. *arXiv preprint arXiv:2004.07737*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Bolton, B. K., Thompson, J., & Thompson, J. L. (2003). *The entrepreneur in focus: Achieve your potential*: Cengage Learning EMEA.

Bundesministerium für Bildung und Forschung (BMBF). (n.d.-a). Cybersecurity research to boost Germany's competitiveness. Retrieved from https://www.bmbf.de/en/cybersecurity-research-to-boost-germany-s-competitiveness-1418.html

Bundesministerium für Bildung und Forschung (BMBF). (n.d.-b). StartUpSecure – Die Initiative für Start-ups in der IT-Sicherheit. Retrieved from https://www.forschung-it-sicherheit-kommunikationssysteme.de/foerderung/startup-secure

Bureau of Labor Statistics. (October 2020). Jobs created by start-ups U.S. 2020. Retrieved from https://www.statista.com/statistics/235515/jobs-created-by-start-ups-in-the-us/

Carland, J. W., Hoy, F., Boulton, W. R., & Carland, J. A. C. (1984). Differentiating entrepreneurs from small business owners: A conceptualization. *Academy of management review, 9*(2), 354-359.

Cavallo, A., Ghezzi, A., & Balocco, R. (2019). Entrepreneurial ecosystem research: present debates and future directions. *International Entrepreneurship and Management Journal, 15*(4), 1291-1321.

Clarysse, B., Wright, M., Bruneel, J., & Mahajan, A. (2014). Creating value in ecosystems: Crossing the chasm between knowledge and business ecosystems. *Research Policy, 43*(7), 1164-1176.

Columbus, L. (2020). Why cybersecurity is really a business problem. *Forbes*. Retrieved from https://www.forbes.com/sites/louiscolumbus/2020/06/25/why-cybersecurity-is-really-a-business-problem/

Cooke, P., Uranga, M. G., & Etxebarria, G. (1997). Regional innovation systems: Institutional and organisational dimensions. *Research Policy, 26*(4-5), 475-491.

Council of Economic Advisors. (2018). The cost of malicious cyber activity to the US economy.

Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining cybersecurity. *Technology Innovation Management Review, 4*(10).

cyberwatching.eu. (2018). *European Cybersecurity and Privacy Research & Innovation Ecosystem*. Retrieved from

Dahlqvist, F., Patel, M., Rajko, A., & Shulman, J. (2019). Growing opportunities in the Internet of Things. *McKinsey Insights*.

Daunfeldt, S.-O., Elert, N., & Johansson, D. (2014). The economic contribution of high-growth firms: Do policy implications depend on the choice of growth indicator? *Journal of Industry, Competition and Trade, 14*(3), 337-365.

de Bruijn, H., & Janssen, M. (2017). Building cybersecurity awareness: The need for evidence-based framing strategies. *Government Information Quarterly, 34*(1), 1-7.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dreißigsacker, A., von Skarczinski, B., & Wollinger, G. R. (2020). *Cyberangriffe gegen Unternehmen in Deutschland*. Retrieved from https://www.pwc.de/de/cyber-security/cyberangriffe-gegen-unternehmen-in-deutschland.pdf

European Commission. (2018). EU data protection rules. Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en

Fagella, D. (2020). The Self-Driving Car Timeline–Predictions from the Top 11 Global Automakers. *by Emerj*.

Federal Ministry for Economics Affairs and Energy. (n.d.). Start-ups: A driving force for growth and competition. Retrieved from https://www.bmwi.de/Redaktion/EN/Dossier/start-ups.html

Feld, B. (2012). Startup communities: building an entrepreneurial ecosystem in your city.

Forum, W. E. (2013). Entrepreneurial Ecosystems Around the Globe and Company Growth Dynamics. *Davos: World Economic Forum*.

Freeman, J., & Engel, J. S. (2007). Models of innovation: Startups and mature corporations. *California Management Review, 50*(1), 94-119.

Friedlmaier, M., Tumasjan, A., & Welpe, I. M. (2018). *Disrupting industries with blockchain: The industry, venture capital funding, and regional distribution of blockchain ventures.* Paper presented at the Venture Capital Funding, and Regional Distribution of Blockchain Ventures (September 22, 2017). Proceedings of the 51st Annual Hawaii International Conference on System Sciences (HICSS).

Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics, 5*, 529-542.

Gálvez, R. H. (2017). Assessing author self-citation as a mechanism of relevant knowledge diffusion. *Scientometrics, 111*(3), 1801-1812.

Gartner. (2019). *Cloud Adoption: Where Does Your Country Rank?* Retrieved from Gartner: https://www.gartner.com/smarterwithgartner/cloud-adoption-where-does-your-country-rank/#:~:text=Since%202015%2C%20the%20U.S.%20has,to%20seven%20or%20more%20years.

Gartner. (2020a). Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 18% in 2021 [Press release]. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2020-11-17-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-18-percent-in-2021

Gartner. (2020b). Gartner Forecasts Worldwide Security and Risk Management Spending Growth to Slow but Remain Positive in 2020 [Press release]. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2020-06-17-gartner-forecasts-worldwide-security-and-risk-managem

Germany Trade and Invest. (2019). *Software and Cybersecurity Market in Germany*. Germany Trade and Invest, Gesellschaft für Außenwirtschaft und Standortmarketing mbh Retrieved from https://www.gtai.de/resource/blob/64552/7f49d0eae50f138fcf5763ecac23c12e/fact-sheet-software-cybersecurity-en-data.pdf

Grand View Research. (2020). *Cyber Security Market Size, Share & Trends Analysis Report By Component, By Security Type, By Solution, By Service, By Deployment, By Organization, By Application, By Region, And Segment Forecasts, 2020 - 2027*. Retrieved from https://www.grandviewresearch.com/industry-analysis/cyber-security-market

Grilli, L., Mrkajic, B., & Latifi, G. (2018). Venture capital in Europe: social capital, formal institutions and mediation effects. *Small Business Economics, 51*(2), 393-410.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software, 10*(1), 1-20.

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., . . . Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals, 13*(2), 586-632.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485-585): Springer.

Helms, M. M., & Nixon, J. (2010). Exploring SWOT analysis–where are we now? *Journal of strategy and management*.

Henrekson, M., & Sanandaji, T. (2014). Small business activity does not measure entrepreneurship. *Proceedings of the National Academy of Sciences, 111*(5), 1760-1765.

Hilpert, M., Meermann, C., & von Borries, J. (2019). *Future made in Germany - Corporate Innovation Benchmark*. Retrieved from Future Made in Germany: https://www.futuremadeingermany.de/

Hogan, M. D., & Newton, E. M. (2015). *Supplemental Information for the Interagency Report on Strategic US Government Engagement in International Standardization to Achieve US Objectives for Cybersecurity*. Retrieved from

Hu, D. J. (2009). Latent dirichlet allocation for text, images, and music. *University of California, San Diego. Retrieved April, 26*, 2013.

Isenberg, D. (2010). How to start an entrepreneurial revolution. *Harvard business review, 88*(6), 41-49.

Isenberg, D. (2011). Introducing the entrepreneurship ecosystem: Four defining characteristics. *Forbes, 25*, 2011.

Ismail, N. (2018). Cloud security – who should take ownership in the enterprise? Retrieved from https://www.information-age.com/cloud-security-ownership-enterprise-123473398/

Jackson, D. J. (2011). What is an innovation ecosystem. *National Science Foundation, 1*(2).

Jathanna, R., & Jagli, D. (2017). Cloud computing and security issues. *International Journal of Engineering Research and Applications, 7*(6), 31-38.

Jaycocks, A. (2019). *Climate Finance and Green Bond Evolution.* PARDEE RAND GRADUATE SCHOOL,

Jevtic, S. (2020). NLP techniques: Semi-supervised topic modelling. Retrieved from https://faculty.ai/blog/nlp-techniques-semi-supervised-topic-modelling/

Kantor, S., & Whalley, A. (2014). Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks. *The Review of Economics and Statistics, 96*(1), 171-188. doi:10.1162/REST_a_00357

Kerr, S. P., Kerr, W. R., & Xu, T. (2017). *Personality traits of entrepreneurs: A review of recent literature* (0898-2937). Retrieved from

Khalid, A. (2017). Why Israeli Cybersecurity Firms Are Moving From Tel Aviv To Boston. Retrieved from https://www.wbur.org/bostonomix/2017/12/18/israeli-cybersecurity-boston

KPMG. (2020). *Venture Pulse Q3, 2020 - Global analysis of venture funding*. Retrieved from https://home.kpmg/xx/en/home/insights.html

Lakshmanan, S. (2019). How, When, and Why Should You Normalize / Standardize / Rescale Your Data? Retrieved from https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Leitersdorf, Y., & Schreiber, O. (2020). Israel's cybersecurity startup scene spawned new entrants in 2019. Retrieved from https://techcrunch.com/2020/01/20/israels-cybersecurity-startup-scene-spawns-new-entrants-in-2019/

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). *Topical word embeddings.* Paper presented at the Twenty-ninth AAAI conference on artificial intelligence.

Liu, Z. (2013). *High performance latent dirichlet allocation for text mining.* Brunel University School of Engineering and Design PhD Theses,

Mansfield-Devine, S. (2019). The state of operational technology security. *Network security, 2019*(10), 9-13.

Mars, M. M., Bronstein, J. L., & Lusch, R. F. (2012). The value of a metaphor: Organizations and ecosystems. *Organizational Dynamics, 41*(4), 271-280.

Mason, C., & Brown, R. (2014). Entrepreneurial ecosystems and growth oriented entrepreneurship. *Final report to OECD, Paris, 30*(1), 77-102.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Minevich, M. (2020). How The US Can Learn About Successful Innovation Strategies From Israel, The Startup Nation. Retrieved from https://www.forbes.com/sites/markminevich/2020/05/29/how-the-us-can-learn-about-successful-innovation-strategies-from-israel-the-startup-nation/

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Moore, J. F. (1993). Predators and prey: a new ecology of competition. *Harvard business review, 71*(3), 75-86.

Mordor Intelligence LLP. (2020). *Digital Forensics Market - Growth, Trends, Forecasts (2020 - 2025)*. Retrieved from https://www.reportlinker.com/p05893116/Digital-Forensics-Market-Growth-Trends-Forecasts.html?utm_source=GNW

Mulcahy, D., Weeks, B., & Bradley, H. S. (2012). We have met the enemy… and he is us: lessons from twenty years of the Kauffman Foundation's investments in venture capital funds and the triumph of hope over experience. *Available at SSRN 2053258*.

Murphy, L. M., & Edwards, P. L. (2003). *Bridging the valley of death: Transitioning from public to private sector financing*: National Renewable Energy Laboratory Golden, CO.

Naujokaitytė, G. (2020). EU finalising plans for pan European cybersecurity centre. *Science | Business*. Retrieved from https://sciencebusiness.net/framework-programmes/news/eu-finalising-plans-pan-european-cybersecurity-centre

Nicholson, C. (n.a.). A Beginner's Guide to Word2Vec and Neural Word Embeddings. Retrieved from https://wiki.pathmind.com/word2vec#:~:text=Given%20enough%20data%2C%20usage%20and,and%20classify%20them%20by%20topic.

Nicotra, M., Romano, M., Del Giudice, M., & Schillaci, C. E. (2018). The causal relation between entrepreneurial ecosystem and productive entrepreneurship: A measurement framework. *The Journal of Technology Transfer, 43*(3), 640-673.

OECD. (2020). *Venture Capital Investments*. Retrieved from: https://stats.oecd.org/Index.aspx?DataSetCode=VC_INVEST#

Onetti, A. (2019). Turning open innovation into practice: trends in European corporates. *Journal of Business Strategy*. doi:10.1108/JBS-07-2019-0138

Pancholi, S., & Strobl, G. (2019). *Digital transformation and its impact on cybersecurity*. Retrieved from https://www.rsm.global/catch-22/consequences-gdpr-cybersecurity#:~:text=It%20is%20no%20surprise%20that,giving%20cybercrime%20more%20tangible%20consequences.

Pierre Audoin Consultants. (2013). *Competitive analysis of the UK cyber security sector*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/259500/bis-13-1231-competitive-analysis-of-the-uk-cyber-security-sector.pdf

Porter, M. E. (1998). *Clusters and the new economics of competition* (Vol. 76): Harvard Business Review Boston.

Press, G. (2017). 6 Reasons Israel Became A Cybersecurity Powerhouse Leading The $82 Billion Industry. Retrieved from https://www.forbes.com/sites/gilpress/2017/07/18/6-reasons-israel-became-a-cybersecurity-powerhouse-leading-the-82-billion-industry/

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*.

Raska, M. (2015). *Confronting cybersecurity challenges: Israel's evolving cyber defence strategy*. Retrieved from http://www.jstor.org/stable/resrep05855

Rice, M. P., Fetters, M. L., & Greene, P. G. (2014). University-based entrepreneurship ecosystems: a global study of six educational institutions. *International Journal of Entrepreneurship and Innovation Management, 18*(5-6), 481-501.

Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the space of topic coherence measures.* Paper presented at the Proceedings of the eighth ACM international conference on Web search and data mining.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*.

Roundy, P. T., Bradshaw, M., & Brockman, B. K. (2018). The emergence of entrepreneurial ecosystems: A complex adaptive systems approach. *Journal of Business Research, 86*, 1-10. doi:10.1016/j.jbusres.2018.01.032

Schumpeter, J. A. (1934). *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle*: Harvard University Press.

Stam, E., Bosma, N., Van Witteloostuijn, A., De Jong, J., Bogaert, S., Edwards, N., & Jaspers, F. (2012). Ambitious entrepreneurship. *A Review of the Academic Literature and New Directions for Public Policy, Report for the Advisory Council for Science and Technology Policy (AWT) and the Flemish Council for Science and Innovation (VRWI)*.

Stam, E., & van de Ven, A. (2019). Entrepreneurial ecosystem elements. *Small Business Economics*, 1-24.

Stam, F., & Spigel, B. (2016). Entrepreneurial ecosystems. *USE Discussion paper series, 16*(13).

Stangler, D., & Bell-Masterson, J. (2015). Measuring an entrepreneurial ecosystem. *Kauffman Foundation. Online at: https://papers.ssrn.com/sol3/papers.cfm*.

Start-up Nation Central. (2017). *Israel: A global center for cyber security*. Retrieved from https://www.startupnationcentral.org/sector/cybersecurity/

StartupBlink. (2020). *Startup Ecosystem Rankings 2020*. Retrieved from https://report.startupblink.com/

Statista estimates. (2019). Total value of venture capital investments per capita in selected countries worldwide as of 3rd quarter 2019 (in U.S dollars). In: In Statista.

Stewart, W. H., Watson, W. E., Carland, J. C., & Carland, J. W. (1999). A proclivity for entrepreneurship: A comparison of entrepreneurs, small business owners, and corporate managers. *Journal of Business Venturing, 14*(2), 189-214. doi:https://doi.org/10.1016/S0883-9026(97)00070-0

Stölzel, T. (2020). Die Kritik an der Fraunhofer-Gesellschaft wird noch lauter. *WirtschaftsWoche*. Retrieved from https://www.wiwo.de/erfolg/gruender/abgeschreckte-investoren-die-kritik-an-der-fraunhofer-gesellschaft-wird-noch-lauter/26598028.html

Tabansky, L., & Ben Israel, I. (2015). The National Innovation Ecosystem of Israel. In *Cybersecurity in Israel* (pp. 15-30). Cham: Springer International Publishing.

U.S. Department of Homeland Security. (2018). *Cybersecurity Strategy*.

Vaugan, A. (2020). Cybersecurity Made in Europe - Mapping the European Cybersecurity Scale-Up Ecosystem V 1.0. Retrieved from https://european-champions.org/de/2020/07/28/cybersecurity-made-in-europe-mapping-the-european-cybersecurity-scale-up-ecosystem/

Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *computers & security, 38*, 97-102.

Weaver, W. (1955). Translation. *Machine translation of languages, 14*(15-23), 10.

Weiblen, T., & Chesbrough, H. W. (2015). Engaging with startups to enhance corporate innovation. *California Management Review, 57*(2), 66-90. doi:10.1525/cmr.2015.57.2.66

Wijngaarde, Y. (2020). Dependency on foreign investors could become a problem for (German) startups. Retrieved from https://blog.dealroom.co/dependency-on-foreign-investors-could-become-a-problem-for-german-start-ups/

Wong, P. K., Ho, Y. P., & Autio, E. (2005). Entrepreneurship, innovation and economic growth: Evidence from GEM data. *Small Business Economics, 24*(3), 335-350.

World Economic Forum. (2016). *Top 10 Emerging Technologies of 2016*. Retrieved from http://www3.weforum.org/docs/GAC16_Top10_Emerging_Technologies_2016_report.pdf

Zhang, C. (2016). *A study on cybersecurity start-ups: A financial approach to analyze industry trends, entrepreneurship ecosystems and start-up exits.* Massachusetts Institute of Technology,

Appendix A

Total number of startups per country

	Number of Startups
United States of America	1969
Israel	476
United Kingdom	343
India	193
Canada	154
France	103
Germany	102
Netherlands	100
China	94
Australia	82
Spain	75
Switzerland	64
Japan	61
Singapore	61
Brazil	52
Sweden	45
Ireland	34
Norway	33
Italy	32
Estonia	29
Poland	26

Denmark	26
Finland	24
South Korea	23
Belgium	21
New Zealand	20
Turkey	19
Russia	18
United Arab Emirates	16
Czech Republic	15
Ukraine	15
Romania	13
Bangladesh	11
Austria	10
Hong Kong	10
Egypt	9
Portugal	9
Lithuania	9
South Africa	9
Indonesia	8
Cyprus	8
Taiwan	8

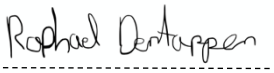
Pakistan	7
Philippines	6
Bulgaria	6
Nigeria	6
Slovakia	6
Vietnam	6
Hungary	5
Mexico	5
Colombia	4
Iceland	4
Kenya	4
Thailand	4
Luxembourg	4
Chile	4
Belarus	4
Jordan	3
Malaysia	3
Greece	3
Nepal	3
Iran	3
Cayman Islands	3

Latvia	2
Tunisia	2
Malta	2
Moldova	2
Uruguay	2
Cameroon	2
Morocco	1
Puerto Rico	1
Bahrain	1
Kuwait	1
Jamaica	1
Serbia	1
Malawi	1
Maldives	1
Armenia	1
Dominica	1
Rwanda	1
Gibraltar	1
Saudi Arabia	1
Argentina	1
Tanzania	1

Georgia	1
Lebanon	1
Albania	1
Qatar	1
Azerbaijan	1
Botswana	1
Uzbekistan	1
Seychelles	1
Ghana	1

Declaration of authorship

I hereby declare that I have written this thesis on my own and with no other help than the literature and other supportive material listed in the appendix. Citations of sentences and parts of sentences are declared as such, while other imitations are clearly marked and linked to original sources with regard to extent and intention of the statements made. This thesis has never been handed in to any examination authority before and it is also not yet published.

Derstappen, Raphael	Siershahn, 19.01.21	
Last Name, First Name	City, Date	Signature