Journal of Intercultural Management
Vol. **13** | No. **4** | **December 2021**
pp. **4–33**
DOI **10.2478/joim-2021-0070**

Márton Gosztonyi
Office for Entrepreneurship Development,
Budapest Business School University
of Applied Sciences, Budapest, Hungary
gosztonyi.marton@uni-bge.hu
ORCID ID: 0000-0003-1887-4913

Comparative Research of Central and Eastern European Startup Researches Based on Artificial Intelligence-Based Natural Language Processing

Received: 03-12-2021; Accepted: 28-12-2021; Published: 13-05-2022

ABSTRACT

Objective: In our study, we analyze Central and Eastern European (CEE) scientific papers published in peer-reviewed scientific journals between 2015 and 2021. We examine what category systems and methods are used in Central and Eastern European start-up researches in the recent years.

Methodology: Our used methodology was structured literature review analysis and artificial intelligence-based natural language processing which is one of the most evolving methodological directions in economics and social sciences at present but it is very rarely used in review analysis of startup research.

Value Added: The NLP method has not been widely used for the analysis of the startup literature. Furthermore, our study is the first which analyzes CEE startups research with NLP technique.

Findings: Based on our results, it can be stated that CEE startup researches follow the big global startup research narratives. However, a specific conceptual network is also emerging which contains several shifts of emphasis compared to the directions of global research.

Key words: startup research, natural language processing, text mining, artificial intelligence, Central and Eastern Europe

JEL codes: M13, C45, M20, I23

Concept of startups

The startup form is one of the manifestations of entrepreneurship (Wennekers & Thurik, 1999; Zacharakis, Reynolds, & Bygrave, 1999) which contributes greatly to the innovative and competitive advantages and economic growth of

a given economy (Praag & Versloot, 2007). As a result, it is no coincidence that the topic is accompanied by strong academic interest, as huge research has been done over the past 80 years on the matter. In the field of startup research, even though knowledge production has been accelerating at a tremendous rate recently, it remains fragmented and interdisciplinary to this day. Therefore, a review of this literature as a research method, which can be described as a systematic way of gathering and synthesizing previous research, is more relevant than ever (Baumeister & Leary, 1997; Tranfield, Denyer, & Smart, 2003). In this sense, a structured literature review (SLR) can be explained as a research method and process for identifying and critically evaluating relevant researches and for collecting and analysing data from such research (Liberati et al., 2009). Ongoing analyses on the startup topic give little attention to the literature of startup researches of Central and Eastern Europe (CEE) with semi-peripheral economies. Consequently, our study provides a comprehensive summary and analysis of CEE research on startup research as a gap-filling article by identifying the key topics and trends.

The definition of the concept of startup itself can be approached from a number of theoretical foundations. Due to this, its definition is spread over a wide spectrum in the literature, highlighting its different mechanisms of action. The concept appeared in the late 60s and early 70s under the collective concept of 'fast-growing businesses'. At that time, theorists focused on the differences in the accumulation of investment and start-up starting costs in their definitions (Aschmann, 1970; Ray, Villeneuve, & Roberge, 1974) as well as the limited funding opportunities that any fast-growing firm would face (Schmidt & Lippitt, 1967). By the 1980s, the concept of a startup was narrowed down to companies operating in a particular industry and so, it mostly referred to companies operating in the semiconductor sector (Angel, 1989; Shoenberger, 1986) which were otherwise known as 'fast-growing electronic startups' (Florida, 2005, p. 256). In parallel with the turbulent growth of the startup sector in the 1990s and 2000s, the definition space expanded, and rapid growth came to the fore in the definition (Saxenian, 1994) in an extremely unstable economic environment (Markusen, 2003). Moreover, new types of financing entities appeared in the definition such as the role of venture capital (Bussgang, 2010).

The role of idea-based operation, informality, and hard work leading to exponential growth in firms (Feld & Mendelson, 2016; Barringer, Jones, & Neubaum, 2005) also played a significant role in the definition of a startup at that time. By the 2000s, startups created new industries and were at the forefront of developing innovative products and services (Fesser & Willard, 1990). Thus, the concept of relevant industry experience and creation was also included in the definition (Hwang & Horowitt, 2012). By 2010, based on feminist economic geography and post-structuralist theories, process – and context-based definitions also prevailed in the definition of startups (Yeung, 2019). Thus, raising the definition of a startup as a narrative (Cockayne, 2019) which is a form of enterprise shows how both corporate action and the way employees work within these are recontextualized. Therefore, the definition on startups started being defined as a kind of working method that promises a better and more modern type of work based on its self-narrative (Marwick, 2013): a love of work (Markusen, 2003; Gill, 2002), a passionate attachment to work (McRobbie, 2002), and a concept of a new urban environment and lifestyle (Florida, 2005). The contextual definition of a startup has also been supplemented with the function of an economic marker, as they mostly appear in countries that are making promising global or regional economic developments driven by technological and knowledge-based forms of enterprise. The concept of a startup has thus also become an economic status symbol of a country, symbolizing the economic sphere populated by highly educated young people at the forefront of business and cultural trends (Marwick, 2013).

The scientific interest of the European semi-peripheral countries turned to startups in the 2000s and then the researches intensified from 2015 albeit with some delays. This coincided with the economic process that the startup ecosystem began to strengthen in the respective countries with it becoming one of the strengthening and sought-after forms of operation of enterprises.

Text mining and artificial intelligence: data and analysis methods

One of the most evolving methodological directions in economics and social sciences at present is the interpretation of texts as data and quantitative text analysis (Sebők, Ring, & Máté, 2021).[1]

Natural language processing (NLP) became one of the leading methodological trends with the Big Data revolution when diverse, non-linear, large-scale, heterogeneous databases appeared (Gosztonyi, 2021). As in the case of the complex methodologies brought to life by Big Data, inductive cognition direction plays a key role in automated text analysis, as opposed to the deductive theoretical directions, so it is especially suitable for exploratory research. However, there are several limitations to the method as Grimmer and Steward (2013) stated in their study. The hermeneutical and phenomenological interpretation of human language cannot currently be part of NLP but with this method, a huge amount of textual data can be analysed and processed. Thus, quantitative text analysis is not intended to replace the analysis of the scientific cognitive in causal analysis but rather to embed the process in an interaction with it.

The use of text mining and NLP is currently in its infancy with regard to the study of startups. However, several very promising studies have been conducted in recent years, such as the research of Glupker et al. (2019) who extended and improved the traditional econometric approaches and used the NLP methodology to analyse how the network position of an investor determines the success of investment in startups. Antretter et al. (2019) identified which startup companies could predict economic failure with the help of the text mining methodology. Santana et al. (2017) used NLP techniques to identify potential, high-value investments in the Crunchbase system. This research was complemented by Xu, Chen, & Zhao (2017) who also analysed the Crunchbase database using the NLP method with the aim of narrowing

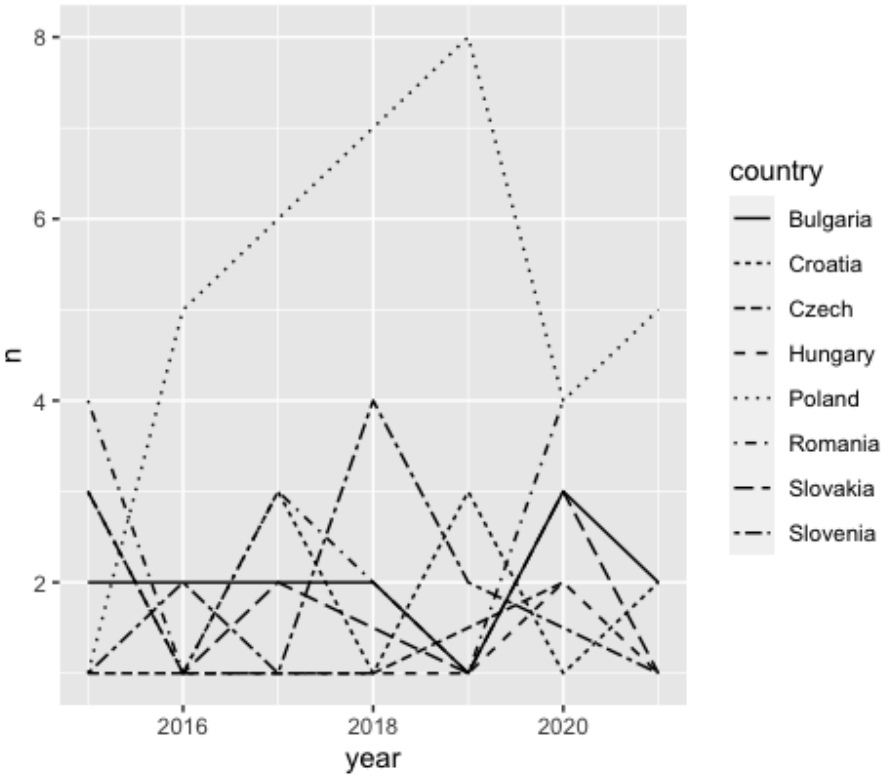
1 This methodological direction can be found under number of names in the literature; text mining, automated content analysis, automated text analysis, natural language processing (NLP).

the list of portfolio companies that venture capitalists can take into account. Dellermann et al. (2017) combined machine learning with traditional qualitative analysis techniques to create a method of the so-called hybrid intelligence for the analysis of a startup's success.

From all this, however, it can be seen that in startup researches, the papers using NLP techniques mostly focus on certain industries, firms, activities, or regions which greatly limits the generalizability of NLP research (Demil et al., 2015; Hermes, Böhm, & Krcmar, 2019). The method has not been widely used for the analysis of the startup literature as well. Furthermore, NLP research on CEE startups is not known either which is why we believe that our research can shed new light on the scientific discourse on these startups.

To identify relevant scientific publications on startups, we used a structured literature review (SLR) following the suggestions of Webster and Watson (2002) in which we (i) searched the leading scientific journal databases based on the given keywords, (ii) selected publications with the appropriate criteria, (iii) did a quick review of the identified publications by reading their titles, abstracts, and the full text and finally (iv) used an NLP analysis to complete the analysis of the selected texts. Our corpus thus consists of scientific texts from eight CEE member states of the European Union (Bulgaria, Croatia, the Czech Republic, Hungary, Poland, Romania, Slovakia, and Slovenia). In our study, we examined scientific works published in English in a peer-reviewed scientific journals between 2015 and 2021. We compiled our database using three databases: Web of Science, EBSCO, and Google Scholar. These databases allowed us to include not only articles published in highly indexed journals but also articles published in journals with smaller global scientific impact. As a result, our entire corpus consisted of 157 articles. However, due to copyright issues, we were only able to analyse 104 articles with text mining. In this regard, our analysis was not based on the entire corpus; it can be interpreted as a quasi-sample study. The distribution of the corpus by country and time interval is summarized in figure 1.

Figure 1. Distribution of corpus documents by country and year (N = 104)



Source: own elaboration.

Figure 1 shows that even though startup articles have a very different number of items per country, their number of appearances is fairly balanced each year. The average number of papers per year is 14.85 (std. ± 2.85). Most of the papers were published in 2018 and 2020 (18 pieces). However, this does not differ significantly from the regional annual average. The proportion of appearances related to the topic varies greatly from country to country. The largest number of published papers are from Poland (28.8%), followed by Romania (14.4%) and

Croatia (13.4%) while the smallest number of scientific publications on the topic is in the Czech Republic (5.8%), and Hungary (5.8%).[2]

The next methodological step was the detailed reading of the full text of the selected publications and their categorization according to their purpose and content in order to produce descriptive statistics from the corpus. After categorizing the sample by author, year, methodology, focus, and topic, the documents were tokenized and lemmatized. In our analysis, we used stopword filtering; on one hand, we used the English stopword lexicon of the Quanteda R program package (Benoit et al., 2018) while on the other hand, we built a specific stopword lexicon on the corpus. After the stopword normalisation, our corpus consisted of 118,882 lemmas.

The first NLP methodology that we used was the term frequency-inverse document frequency (TF-IDF) analysis which calculates the inverse value of the word-document frequency (Aizawa, 2003). Based on this, it can be concluded that if a given word is associated with a high TF-IDF score, it occurs frequently in the document while in the whole corpus it is rare. The inverse value of TF-IDF, therefore, shows how significant the given expression is in general in its topic. In our study, the TF-IDF index was calculated and normalized according to the following formula:

$$f_{\text{term}} = \sum_{i}^{n} \text{sentence\_occurrences}_{i}$$

where F_{term} represents the frequency of the expression, i covers the number of documents, and sentence_occurrencesesi is the number of sentences in the i document in which the expression is found. Normalized TF-IDF values were calculated using the following formula:

$$Z_{\text{term}} = \frac{f_{\text{term}} - min(F)}{max(F) - min(F)} * 99 + 1$$

where Z_{term} is the frequency index of an expression, $min(F)$ is the minimum value of the frequency of the expression while $\max(F)$ is the maximum value of the frequency of the expression (Kuzminov et al., 2018). With descriptive

2 We have created a user-friendly online application for descriptive statistics, which is available at the following link: https://startupbge.shinyapps.io/Startuppapers.

statistics such as term frequency (TF) and TF-IDF the corpus elements were examined to reveal what defining terms appear in the text. However, we performed several additional NLP analyses on the startup corpus in order to run our results and models through multi-segment validation.

By performing sentiment analysis, our goal was to explore the latent dimensions of the texts (Laver, Benoit, & Garry, 2003), i.e., to extract information from the content of individual texts that express evaluation and then examine their change over time (Liu, 2010). In our research, we performed an emotion analysis as well in which we placed the lemmas not only on a positive-negative-neutral scale but also on an emotional scale. Thus, in the sentiment analysis, we used an n-point scale into which words were entered using a lexicon-based, automatic classification (Sebastiani, 2002; Prabowo & Thelwall, 2009). For this, we used Saif and Turney's (2013) National Research Council of Canada Emotion Lexicon (NRC) which is an emotional scale-based dictionary of 10 elements (positive, negative, anger, expectation, disgust, fear, joy, sadness, surprise, and trust). Thus, in our emotion analysis, we performed a predefined dictionary-based analysis based on the frequency of keywords within a given category (Young & Soroka, 2012; Grimmer & Stewart, 2013). The method has also made it possible to capture the emotions of texts (Hatzivassiloglou & McKeown, 1997; Watanabe, 2021; Rudkowsky et al., 2018) and to analyse subjective emotion associations (Kim & Hovy, 2004; Wilson, Wiebe, & Hoffmann, 2005).

Recently, much attention has also been paid to the generative probabilistic models of text corpus which aim to identify data representations that reduce the length of the description and reveal statistical structures between or within documents. Thus, in the further analysis of the corpus, we used the method of word embedding based on artificial intelligence and unsupervised machine learning. The method of word embedding is constructed through neural networks, where the network produces a vector representation of each word during the learning process. For our word embedding analysis, we used a 300-dimensional vector which is common in the literature (Sebők, Ring, & Máté, 2021). The distance can be used to determine the semantic relationship between each lemma. Expressions with semantically similar content are close to each other while different ones are located far apart in the multidimensional space. Thus, word embedding

is basically a technique of a dense vector representation of words in a corpus where the dimension of the word vector is smaller than the size of the corpus itself. The vectors created during the analysis, in contrast to the results of co-occurrence analysis for example, capture semantic relationships between the linguistic elements examined, as the distribution hypothesis suggests that semantically similar words generally have a similar contextual distribution (Harris, 1954).

Finally, the corpus was also analysed using text scaling to examine how well the thematic concepts defined the documents. In text scaling, we used the wordfish method which can be classified as an unsupervised machine learning methodology (Slapin & Proksch, 2008). Text scaling approaches are primarily used to directly identify the latent positions of political actors (Laver, Benoit, & Garry, 2003; Proksch & Slapin, 2010). However, they have also proved to be an important method of analysis in our literature review as well. Unlike dictionary methods, wordfish does not work with pre-recorded reference points but explores expressions that distinguish texts from each other. The method based on IRT (item response theory) which, in turn, is based on the fact that the concepts move in a small dimensional space can be described by the parameter θ1 of the concept i. The position of a text in this space influences the use of words in the same (Hjorth et al., 2015; Grimmer & Stewart, 2013). Wordfish assumes that the words in the documents follow a Poisson distribution. More specifically, wordfish is a version of an ideal point model of Poisson where, based on a collection of documents and the frequency of the jth word in the ith document, Wijs is derived from a Poisson distribution of λij, which can be modelled by taking into account the document length (αi), the lemma frequency (ψj), the level to which the lemma identifies the direction of the underlying ideological space (βj), and the underlying position of the document (θi) (Slapin & Proksch, 2008; Sebastiani, 2002). According to Slapin and Proksch (2008), the method can be summarized in the equation below:

$$\lambda_{ij} = \exp\left(\alpha_i + \psi_j + \beta_j \times \theta_i\right)$$

In summary, our structured literature review is based on hybrid classification NLP techniques which we followed to increase the accuracy of classifications as well as to deepen the multi-segment understanding of the corpus.

Research questions and hypotheses

Our study analyses Central and Eastern European startup researches published in English in scientific journals between 2015 and 2021. Consequently, in our exploratory research, we looked for the answer pertaining to what extent the research in different countries shows similarity or to what extent the concept directions of research have changed during the seven-year examined time interval. Consequently, our main research hypotheses can be summarized as follows:

H1: Scientifically based startup research in semi-peripheral CEE countries is themed along with similar topics.

H2: Scientifically based startup research in semi-peripheral CEE countries analyses startups in a neutral emotional context.

H3: In the case of startup research in CEEU, which has not developed any separate conceptual or analytical specificity, startup research is interpreted in this region through global scientific concepts and interpretations.

NLP analysis of startup studies

To firstly analyse the descriptive statistics of our corpus, we classified the studies according to several factors so we analyzed the research focus of the articles[3], the main topic, and the methodology used in the researches.

The vast majority of studies (72.4%) focus on a given state, and only a small proportion (27.6%) include comparative research between countries. The distribution of texts by topic (table 1) shows that the texts can be divided into six topic groups, of which about ¼–¼ are texts analysing startup ecosystems (28.6%) and texts presenting startup ideas and business models (24.8%). The share of articles that focus on the topic of finance and management is also outstanding (20.0%).

3 Focus on research within a country or internationally comparative focus.

Table 1. Distribution of corpus documents as topics (N = 104)

	Frequency	Percent
Ecosystem	30	28.6
Idea generation and Model	26	24.8
Finance and Management	21	20.0
Survival and growth	14	13.3
Sustainability	9	8.6
Gender and Culture	5	4.8

Source: own elaboration.

The distribution of research topics by country is presented in table 2. The table shows that there is a very large variance between research topics within countries. Ecosystem and Startup idea and model research topics are those topics we encounter in all countries. The ecosystem has a similar percentage distribution (average percentage: 24.06%, standard deviation: 8.8%) throughout the countries, although idea generation and models show a large variance, as nearly half of Bulgarian (41.67%) and Romanian (46.67%) research belongs to this topic. In the case of any other topic, we always see a country that completely lacks any research on that topic. In other words, it appears in a very primitive manner. Financial and management research, for example, appears largely in Croatia (35.71%), Poland (26.6%), and Romania (26.6%), while in Slovakia and Slovenia, we do not find any research on this topic.

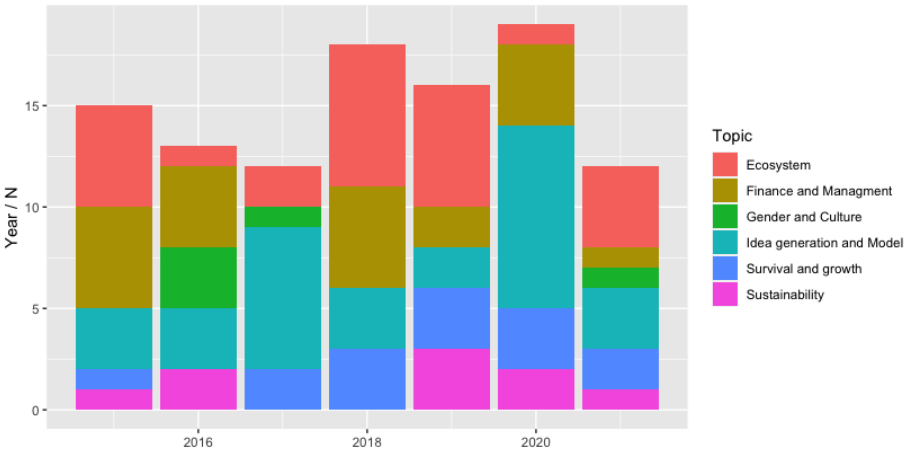
Table 2. Distribution of research topics by country (%, N = 104)

	Bulgaria	Croatia	Czech	Hungary	Poland	Romania	Slovakia	Slovenia
Ecosystem	0.25	0.21	0.17	0.17	0.27	0.13	0.36	0.36
Finance and Management	0.17	0.36	0.17	0.17	0.27	0.27	0	0
Gender and Culture	0	0.07	0	0	0.07	0	0.09	0.09
Idea generation and Model	0.42	0.35	0.33	0.17	0.13	0.47	0.18	0.36
Survival and Growth	0.17	0	0.33	0.17	0.13	0.07	0.18	0.18
Sustainability	0	0	0	0.33	0.13	0.07	0.18	0

Source: own elaboration.

If we look at the yearly breakdown (figure 2), it can be seen that research on a larger proportion of topics (ecosystem, finance and management, survival and growth, idea and model) is written every year (with a different number of papers, of course). However, we can only observe the short-term emergence and disappearance of other topics. Such is the case with sustainability which will disappear completely in 2017 and 2018, or studies in the field of gender and culture which were carried out only in 2016 and 2017 by researchers in the region.

Figure 2. Distribution of corpus text topics by country and year (N = 104)



Source: own elaboration.

From a methodological point of view, only a fraction of the research can be classified as longitudinal research (16.2%). Quantitative analyses accounted for 50.5% of the studies while qualitative methodologies accounted for 47.6% of the research. Therefore, mixed-methodology texts appear only in a relatively small proportion (1.9%). However, within the qualitative and quantitative methodological categories, the methods used in specific studies spread over a very wide spectrum which are summarized in table 3.

Table 3. Distribution of the corpus by research methodology (N = 104)

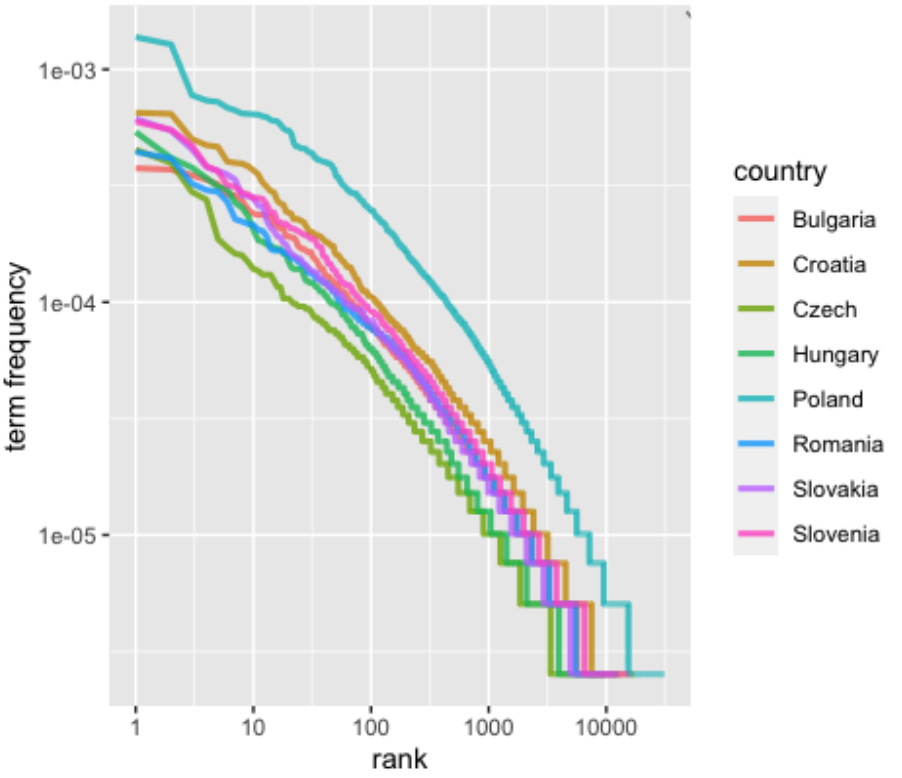
	Frequency	Percent
Survey	51	48.6
Text analysis	24	22.9
Literature review	15	14.3
Case study	12	11.4
Simulation	3	2.9

Source: own elaboration.

Papers using the survey method accounted for almost half of the articles (48.6%) but the proportion of texts using interview methods and focus group methods (22.9%) and texts containing literature reviews (14.3%) were also high. Researchers work with complex system-level analysis methodologies only in a few cases (neural networks, spatial-agent dynamic model, simulation methodology, etc.) due to which, these analyses accounted for only 2.9% of the texts.

Moving from descriptive statistics to text mining methods, one of the most commonly used indicators in our study was the weighed TF-IDF index. Based on the TF-IDF indexes, after a log-log normalization, it became comparable to what extent the texts differ from the average text corpus by country (figure 3).

Figure 3. TF-IDF indexes of documents classified by country in a log-log system (N = 104)

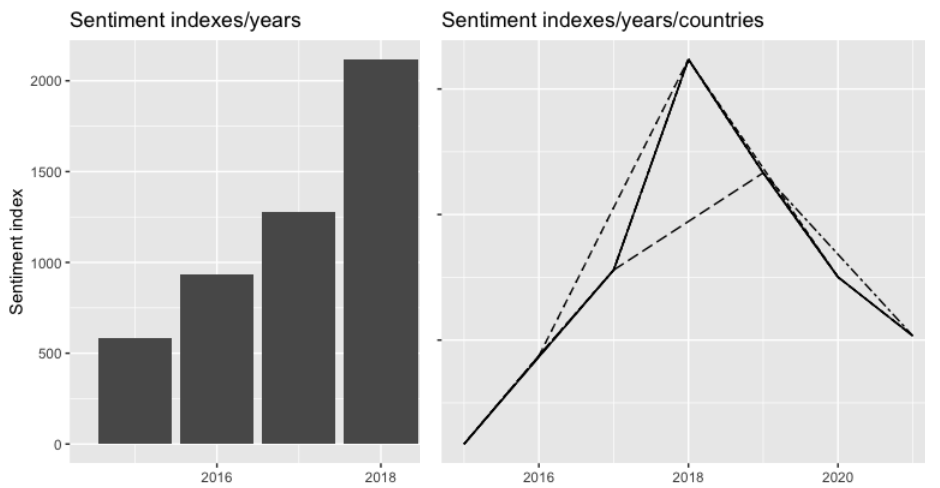


Source: own elaboration.

It can be seen from the figure that, not surprisingly, the TF-IDF indexes of documents follow Zipf's law (Zipf, 1949). At the same time, a more important result is that the TF-IDF values of the countries are strongly correlated with each other. Although the difference based on the largest lemma is found in the case of Poland, the Czech Republic, and Hungary, the wording of the texts shows a high degree of consistency in each country. Based on the lemmas, a strong congruence can be observed for the entrepreneur/enterprise lemma ("entrepreneurship", "entrepreneur", "business"), for the "startup" lemma, for the "innovation" lemma, in the forms of investment ("venture capital", "funds") lemma, and for the "accelerator" lemma. In the case of differentiation in Poland, the "biomass" and "fintech" lemmas cause a large difference, while in the case of Hungary and the Czech Republic, the mention of a specific national startup by name causes the differentiation.

Therefore, in the case of texts, we see a difference in certain topics, but at the same time, a high degree of consistence in terms of lemmas. To delve deeper into the analysis of these narratives and to grasp the emotional attitudes of the research about the topic, we conducted a sentiment analysis. To do this, we first constructed a sentiment index based on negative and positive sentiment values which were aggregated by year and country (figure 4).

Figure 4. Sentiment index of texts by year and country

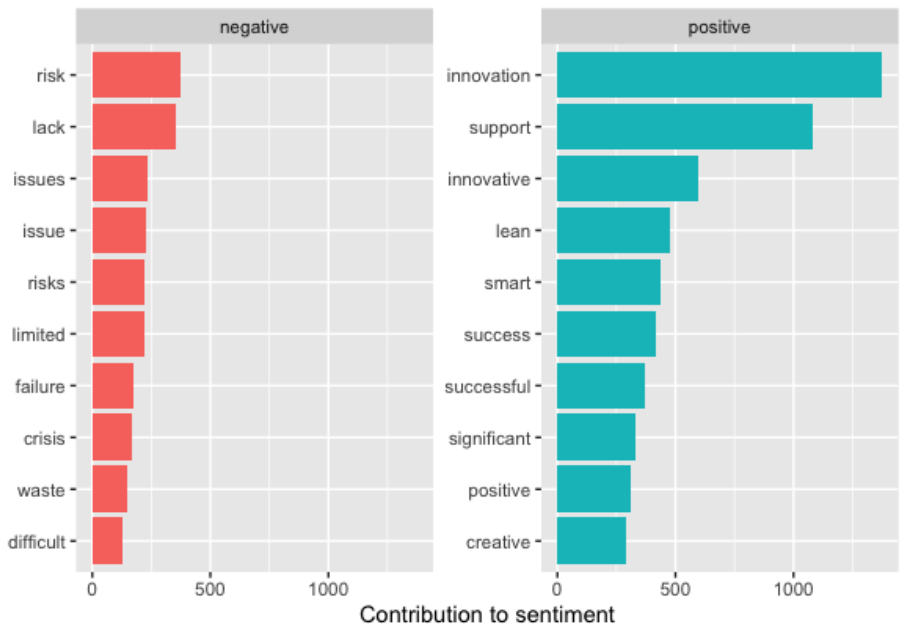


Source: own elaboration.

Figure 4 shows that the texts reflect positive emotional attitudes towards start-ups every year without any exception. We can also see a strong increase in this non-neutral emotional relationship until 2018 after which, there is a slow decline. This positive attitude increasing and then decreasing is true for the papers that are written in any country, although in the case of the Czech Republic, we can observe a faster increase in the positive sentiment index. In the case of Hungary, the highest index score related to the topic is still lower than the regional average.

The most common words with positive and negative sentiment indices in the corpus are shown in figure 5. These sentiment words describe a linguistic context in which the authors place their analyses. Based on these, startups are placed in a narrative space by researchers that is positively associated with "innovation", "support", "successes", and "creativity: while from a negative direction it also has "high risk", "scarcity", "problems", "failure", and "difficulty".[4]

Figure 5. Lemmas with the highest positive and negative sentiment indexes

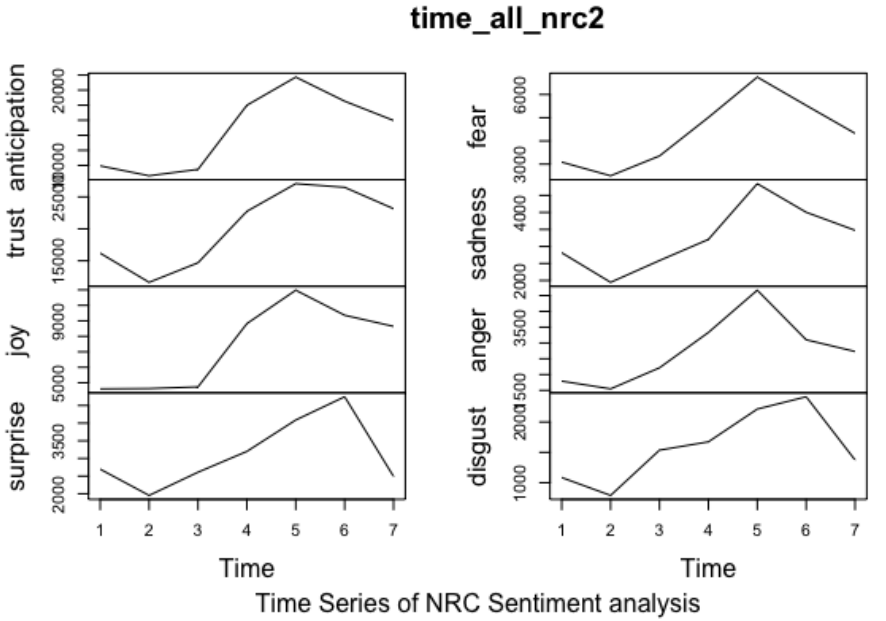


Source: own elaboration.

4 Positive and negative sentiment analysis was also performed by country as well, but we obtained a highly correlated result with the aggregate level analysis.

We captured the positive and negative narrative space by country and year with the emotion analysis as well. The results of the eight-value NRC emotional analysis are summarized in figure 6. The lines show that the clear rise and subsequent decline of each emotion can be observed throughout the text corpus. Based on the emotion analysis, the texts published in 2019 show a peak in both positive ("expectation", "trust", "joy") and negative ("fear", "sadness", "anger") emotions. The year 2020 shows outstanding index numbers for "surprise" and "disgust". Although the year 2019 marks a turning point in emotions, the texts also show that we can be aware of a much steeper rise in negative emotions than in positive emotions.

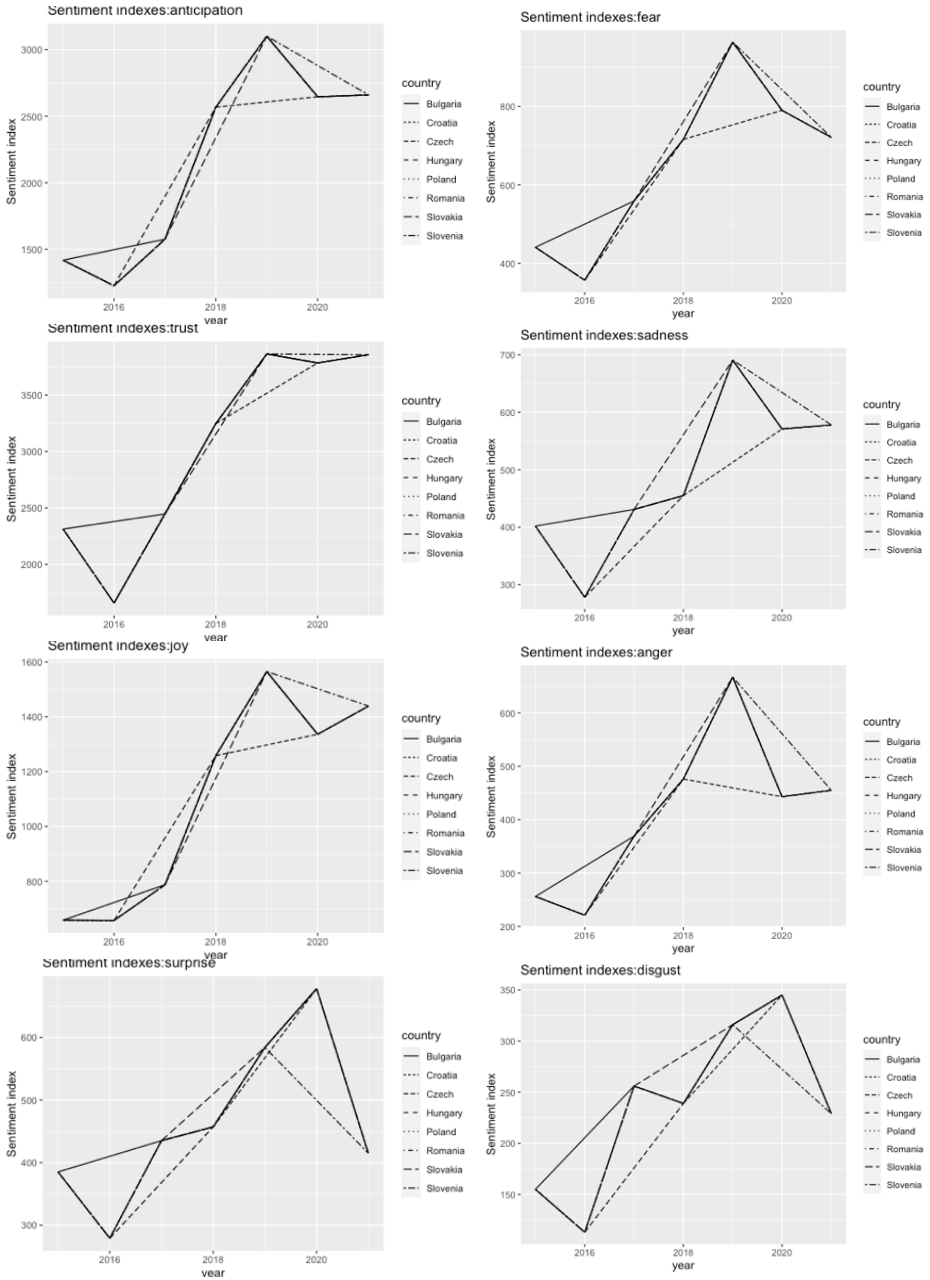
Figure 6. NRC sentiment analysis 2015–2021



Source: own elaboration.

If we break down the emotion indices by country (figure 7), we can see that a fairly high match emerges between the countries in the time in both positive and negative emotions. Exceptions to this are Hungary, where the NRC indices have a later peak and a less high index number, and the Czech Republic and Romania, where steeper increases can be observed in the indices.

Figure 7. NRC sentiment analysis 2015–2021 and broken down by country



Source: own elaboration.

In summary, it can be seen that the "golden year" of the topic in CEE was the year 2018, as the researchers in this year across the region have placed startup studies in a very positive interpretive framework. But this positive attitude seems to be declining throughout the following years. We do not see any major differences from this trend in any countries in the case of positive emotions although in the case of negative emotional words, we see those papers written in Hungary, the Czech Republic, and Romania differ slightly from the trend lines.

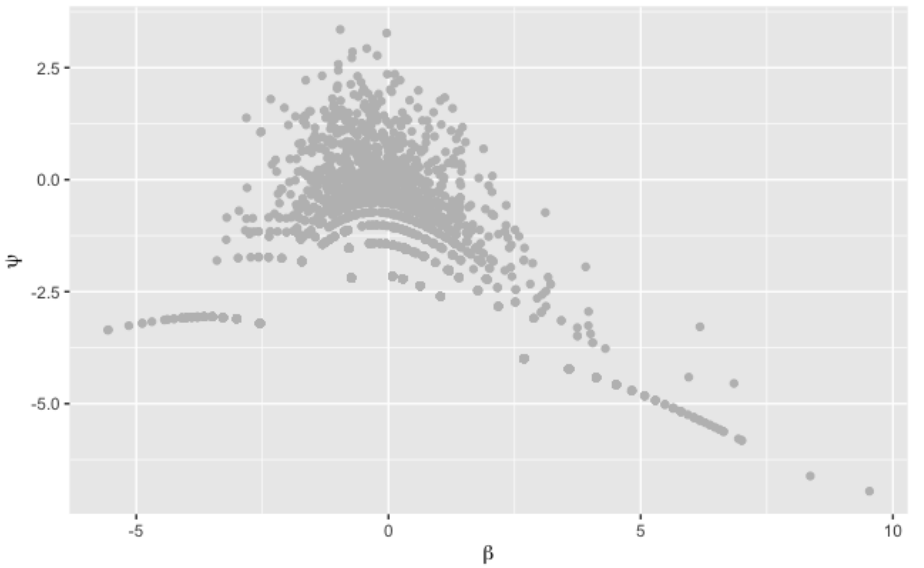
Leaving aside the analysis of sentiment analysis, it is worth seeing what interpretive space the corpus spreads out along concepts, words, and lemmas. Thus, to understand what the web of interpretation is, we need to look at the conceptual and terminology narrative that is used in CEE research. To do this, we analysed the corpus by text scaling firstly. In our analysis, we used β[5] and ψ[6] values[7]. We published our results in the classical "Eiffel Tower" figure which illustrates the frequency of words and their influence on the scale (figure 8). In the "Eiffel Tower" figure, three major categories of word usage can be distinguished: (i) frequently used but neutral words (high ψ, low β), (ii) less frequently used but more determinative words (high β, low ψ), and (iii) moderately used and moderately determinative words (medium β, medium ψ). In the first category, we find words ("visegrad", "eastern", "survey") that characterize the articles mainly territorially and methodologically. In the second category, it is much more difficult to find a common set of interpretations; we find words such as "stakeholders", "events", "experiences", "top down", "collaboration", "model". All of these words appear infrequently and yet, they are significant to the corpus of text. These words and concepts mainly summarize business cooperation models and directions. Finally, in the middle, we find the most commonly used neutral words such as "business", "entrepreneurship", "development", "project", "growth", "financial", "investment", "capital", and "startup".

5 The weight associated with the words which shows the relative importance of the word.

6 Word fixed effects, which corrects a different word frequency.

7 Considering the length of the documents and the word frequency, words with a negative β value are more often used by texts with a negative θ coefficient.

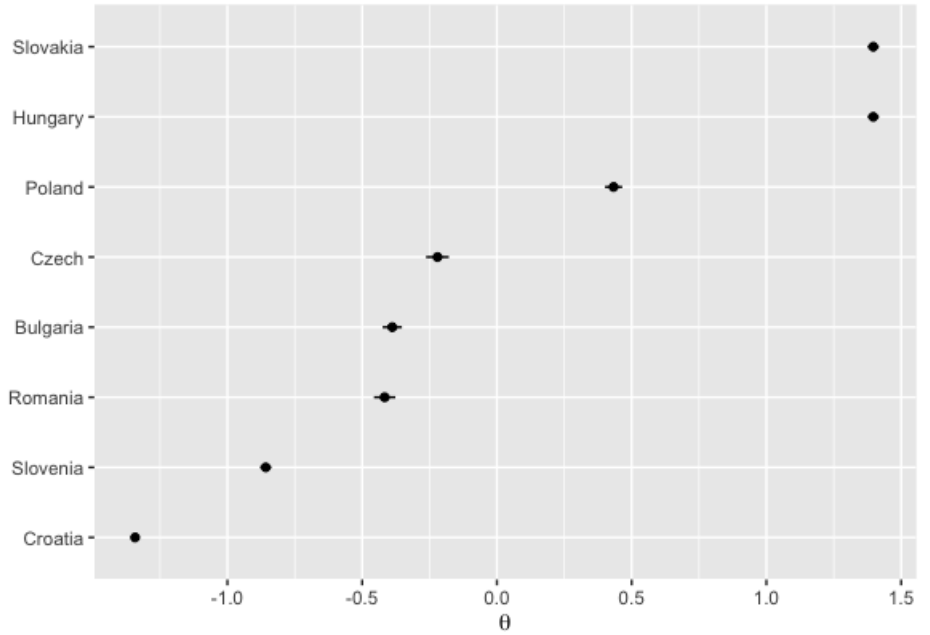
Figure 8. Eiffel Tower illustration of the text scaling



Source: own elaboration.

Comparing the country metadata, the aggregate values of the θ positions for a word are obtained as shown in figure 9. It can be seen that Slovakian and Hungarian documents use term occupies a high relevance, while the other end of the relevance scale is marked by the lemmas of the Slovenian and Croatian studies.

Figure 9. Level of documents based on θ positions



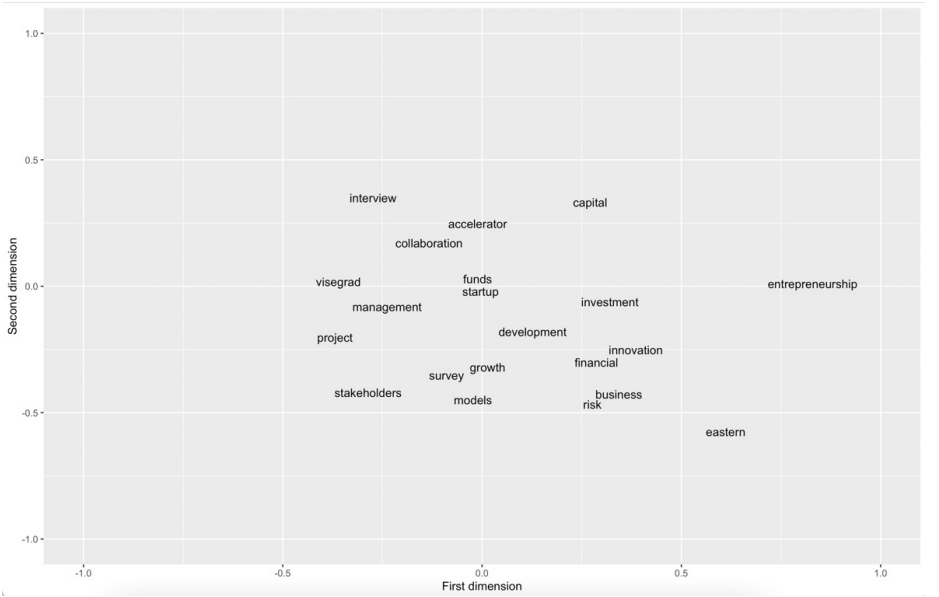
Source: own elaboration.

All this means that the texts strongly represent a regional conceptual network. The role of investors, investments, startup models, and the issue of collaborations appear very strongly in this conceptual regional network. The core elements of the conceptual network focus on entrepreneurship, development, growth, and financial investment. However, it can be seen that this network of interpretation is not uniform across the region. For example, studies in Hungary and Slovakia focus mainly on investments and startup models while research in Slovenia and Croatia focuses more on regional specificities.

The unification and synthesis of the conceptual network were performed by the method of word embedding based on unsupervised artificial intelligence learning. In the semantic analysis of relations, we used the previously explored topics, clusters, and results, so that the stretched narrative space consists of twenty-one lemmas (figure 10). In the multidimensional narrative space, the lemma denoting the twenty-one corpus is separated. Obviously, we do not

aim to sharply delimit the boundaries between relations but it is important to highlight the conceptual attributes of different corpus clusters.

Figure 10. Word embedding-based narrative space of the startup research corpus



Source: own elaboration.

As a result of the analysis, a very interesting circular narrative space emerges, in which, however, clear cluster boundaries are difficult to identify. It can be seen that the center of the conceptual space is occupied by the startup lemma which is naturally a central organizing element of the studies and the other concepts of the corpus emerge around this organizing point in concentric circles. The role of investments ("funds") can be linked the closest to the concept. In the next circle, we find a diverse set of concepts that can still be linked to investments ("financial", "investment", "capital") but they are also closely related to the examination of development directions ("development") and growth ("growth"). The most positive and negative combinations of sentiment analysis ("innovation", "risk") are also in this circle. In this conceptual space, we also find the concepts of collaboration and management ("accelerator", "collaboration")

which are as closely related to the central startup concept as well as the concept of funds, but they apparently represent a completely different approach from that of the growth and model direction. From a methodological point of view, a particularly interesting result is that the quantitative research direction (survey) is more related to the words of business, finance, and development, while the interview methodology is more related to the accelerators and directions examining collaboration. It is important to point out that the two concepts are sharply detached from the narrative center: "entrepreneurship" on the one hand and "eastern" on the other are markedly distinct in the conceptual narrative. All this may mean that in the texts, startups are identified differently from the classical enterprises as well as the use of regional analytical aspects is extremely typical of startup studies in CEE.

As a summary of our research results, we publish the verification of our hypotheses (table 3). We rejected our hypotheses in all cases which means that based on the NLP analysis of the CCE startup research, the differences in the topic choices can be detected and the researches largely frame the topic not neutrally but with positive emotions while the CCE research uses a region-specific conceptual network in their research.

Table 3. Verification of hypotheses

Hypothesis	Verification	Method
H1	Not supported	TF-IDF
H2	Not supported	Sentiment and emotion analysis
H3	Not supported	Text scaling, Word embedding

Source: own elaboration.

Conclusion

In our study, we examined the scientific studies of the semi-peripheral CEE countries, with EU membership published on the topic of startups between 2015 and 2021, using text mining and unsupervised machine learning based on artificial intelligence. In our study, we looked at what descriptive statistics can be used to characterize these papers as well as our capture of the different thematization, emotion, and use of concepts trends in these studies.

Based on our results, it can be concluded that there is little research on startups in CCE in the international discourse. Countries can be characterized with low-average publications in English-language peer-reviewed journals. Researchers carry out researches based on qualitative and quantitative methods in almost the same proportion but very little research is conducted that includes longitudinal and international comparisons. The topics of the researches can be brought into line with the topics appearing in the global research narrative. However, it can also be seen that these topics appear in different proportions along the countries. of the region. The startup ecosystem is a topic that we encounter in almost all countries with a similar publication rate but for other topics, we can observe strong differences across countries. In a global comparison, it is also striking that some important topics (sustainability, gender, culture) are on the periphery of regional research. Based on the NLP study, it can be concluded that the studies work with strongly similar concepts across countries. There is an outstanding match for the entrepreneur/enterprise lemmas, the “startup” lemma, the “innovation” lemma, the lemmas of investment forms, and the development lemma. It follows from all this that we can see differences in the research focuses across countries in terms of topics but we find a high degree of matching in terms of vocabulary as well.

Researches follow a similar trend line in the emotional relationship to the startup topic. The “golden year” of the topic in CEE was the year 2018. In this year, researchers across the region placed startup studies in a very positive interpretive framework. However, this positive attitude seems to be steadily declining thereafter.

Scientific writings in the region also stretch a narrative space around the interpretation of the topic, described in a positive direction by innovation, support, success, and creativity and in a negative direction by high risk, scarcity, problems, failure, and difficulty. All this narrative space also affects the use of the concept which consequently forms a peculiar conceptual network in CCE. Unsurprisingly, the regional focus is very strong in the conceptual web as well as the role of investors, investments, the role of business models, and the issue of cooperation. The core elements of the conceptual network focus on development, growth, and financial investment.

In summary, this means that startup researches in CCE countries with a semi-peripheral economy follow global research directions with a focus on a much less turbulent economic environment. This has created a specific narrative space in research, characterized by a very positive attitude towards the topic as well as a specific CCE conceptual network which follows the global perspectives of interpretation but also includes a number of emphasis shifts.

References

Aizawa, A. (2003). An Information-theoretic Perspective of tf–idf Measures. *Information Processing & Management*, *39*(1), 45–65.

Angel, D. P. (1989). The Labor Market for Engineers in the US Semiconductor Industry. *Economic Geography*, *65*, 99–112.

Antretter, T., Blohm, I., Grichnik, D., & Wincent, J. (2019). Predicting new venture survival: a Twitter – based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights*, *3*(12), 22–33.

Aschmann, H. (1970). The Natural History of a Mine. *Economic Geography*, *46*, 172–189.

Barringer, B. R., Jones, F. F., & Neubaum, D. O. (2005). A Quantitative Content Analysis of the Characteristics of Rapid-Growth Firms and Their Founders. *Journal of Business Venturing*, *20*(5), 663–687.

Baumeister, R. F., & Leary, M. R. (1997). Writing Narrative Literature Reviews. *Review of General Psychology, 1*, 311–320. DOI: 10.1037/1089-2680.1.3.311.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R Rackage for the Quantitative Analysis of Textual Data. *Journal of Open Source Software, 30*(3), 774. DOI: 10.21105/joss.00774.

Bussgang, J. (2010). *Mastering the VC Game*. London: Penguin.

Cockayne, D. (2019). What Is a Startup Firm? A Methodological and Epistemological Investigation into research objects in economic geography. *Geoforum*, *107*, 77–87.

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2017). Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method. *SSRN Electronic Journal*, *12*(2), 22–44, DOI: 10.2139/ssrn.3159123.

Demil, B., Lecocq, X., Ricart, J. E., & Zott, C. (2015). Introduction to the SEJ Special Issue on Business Models: Business Models within the Domain of Strategic Entrepreneurship. *Strategic Entrepreneurship Journal, 9*(1), 1–11.

Feld, B., & Mendelson, J. (2016). *Venture Deals*. Hoboken: Wiley.

Fesser, H. R., & Willard, G. E. (1990). Founding Strategy and Performance: A Comparison of High and Low Growth High Tech Forms. *Strategic Management Journal, 11*(2), 87–98.

Florida, R. (2005). *Cities and the Creative Class*. London: Routledge.

Gill, R. (2002). Cool, Creative, and Egalitarian? Exploring Gender in Project-Based New Media Work in Europe. *Information, Communication & Society*, *5*, 70–89.

Glupker, J., Nair, V., Richman, B., Riener, K., & Sharma, A. (2019). Predicting Investor Success Using Graph Theory and Machine Learning. *Journal of Investment Management, 17*(1), 92–103.

Gosztonyi, M. (2021). A Big data és a részvételiség hatása a tudományos megismerésre és oktatásra. In P. Furkó, É. Szathmári (Eds.), *Tudomány, küldetés, társadalmi szerepvállalás: STUDIA CAROLIENSIA – A Károli Gáspár Református Egyetem 2020-as évkönyve*. Budapest: L'Harmattan Kiadó.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), 267–297.

Harris, Z. S. (1954). *Distributional structure. Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the association for computational linguistics* (pp. 174–181). Madrid, Spain.

Hermes, S., Böhm, M., & Krcmar, H. (2019). Business Model Innovation and Stakeholder Exploring Mechanisms and Outcomes of Value Creation and Destruction. In T. Ludwig, V. Pipek (Eds.), *Proceedings of the 14. Internationale Tagung Wirtschaftsinformatik (WI 2019)*. Siegen, Germany.

Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M. E., & Kurrild-Klitgaard, P. (2015) Computers, Coders, and Voters: Comparing Automated Methods for Estimating Party Positions. *Research & Politics*, *2*(2). DOI: 2053168015580476.

Hwang, V., & Horowitt, G. (2012). *The Rainforest: The Secret to Building the Next Silicon Valley*. Los Altos Hills: Regenwald.

Kim, S.-M., & Hovy, E. (2004, August 23–27). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on computational linguistics (COLING 2004)* (pp. 1367–1373). Geneva, Switzerland.

Kuzminov, I., Bakhtin, P., Khabirova, E., Kotsemir, M., & Lavrynenko, A. (2018). Mapping the Radical Innovations in Food Industry: A Text Mining Study. *Higher School of Economics Research Paper No. WP BRP 80/STI/2018*. DOI: 10.2139/ssrn.3143721.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review 97*(2), 311–331.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine, 151*, W–65. DOI: 10.7326/0003-4819-151-4-200908180-00136.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing, 2*(2010), 627–666.

Markusen, A. (2003). Fuzzy Concepts, Scanty Evidence, Policy Distance: The Cas for Rigour and Policy Relevance in Critical Regional Studies. *Regional Studies*, *37*, 701–717.

Marwick, A. (2013). *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*. New Haven: Yale University Press.

McRobbie, A. (2002). Clubs to Companies: Notes on the Decline of Political Culture in Speeded up Creative Worlds. *Cultural Studies*, *16*, 516–531.

Praag, M., & Versloot, P. H. (2007). What is the Value of Entrepreneurship? A Review of Recent Research. *Small Business Economics, 29,* 351–382. DOI: 10.1007/s11187-007-9074-x.

Prabowo, R., & Thelwall, M. (2009). Sentiment Analysis: A Combined Approach. *Journal of Informetrics*, *3*(2), 143–157.

Ray, M. D., Villeneuve, P. Y., & Roberge, R. A. (1974). Functional Prerequisites, Spatial Diffusion, and Allometric Growth. *Economic Geography*, *50*, 341–351.

Saif, M., & Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada, 2.*

Santana, J., Hoover, R., & Vengadasubbu, M. (2017). Investor Commitment to Serial Entrepreneurs: A Multilayer Network Analysis. *Social Networks*, *48*, 256–269.

Saxenian, A. (1994). *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge: Harvard University Press.

Schmidt, W. H., Lippitt, G. L. (1967). Crises in a Developing Organization. *Harvard Business Review, 45*(6), 102–112.

Schoenberger, E. (1986). *The Cultural Crisis of the Firm*. London: Blackwell.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Sebők, M., Ring, O., & Máté Á. (2021). *Szövegbányászat* és *mesterséges intelligencia R-ben*. Budapest: Typotex Kiadó.

Slapin, J. B., & Proksch, S. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722.

Proksch, S.-O., & Slapin, J. B. (2010). Position Taking in European Parliament Speeches. *British Journal of Political Science, 52* (2010), 587–611.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management, 14,* 207–222. DOI: 10.1111/1467-8551.00375.

Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures, 15*(2), 81–102.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly, 26*(2), xiii–xxiii.

Wennekers, S., & Thurik, R. (1999). Linking Entrepreneurship and Economic Growth. *Small Business Economics, 13*(1), 27–56. DOI: 10.1023/A:1008063200484.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceeding of the conference on empirical methods in natural language processing (EMNLP 2005)* (pp. 347–354). Vancouver, BC, Canada.

Xu, R., Chen, H., & Zhao, J. L. (2017). Predicting Corporate Venture Capital Investment. In *38th International Conference on Information Systems (ICIS 2017): Transforming Society with Digital Innovation*. Republic of Korea: Association for Information Systems.

Yeung, H. W-C. (2019). Rethinking Mechanism and Process in the Geographical Analysis of Uneven Development. *Dialogues in Human Geography, 9*(3), 226–255.

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication, 29*(2), 205–231.

Zacharakis, A., Reynolds, P. D., & Bygrave, W. D. (1999). *National Entrepreneurship Assessment: United States of America. 1999 Executive Report*. Kansas City, Mo.: Kauffman Center for Entrepreneurial Leadership.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley.