Papers in Evolutionary Economic Geography

# 19.20

New(s) data for Entrepreneurship Research? An innovative approach to use Big Data on media coverage

Johannes von Bloh & Tom Broekel & Burcu Özgun & Rolf Sternberg

Utrecht University Human Geography and Planning

New(s) data for Entrepreneurship Research? An innovative approach to use Big Data on media coverage

Johannes von Bloh[a] & Tom Broekel[b] & Burcu Özgun[b,c] & Rolf Sternberg[a]

a) Institute of Economic and Cultural Geography, Leibniz University, Hanover, Germany
b) Department of Human Geography and Spatial Planning, Utrecht University, Utrecht, The Netherlands
c) Department of Economics, Middle East Technical University, Ankara, Turkey

This is a pre-print of an article published in Small Business Economics. The final authenticated version is available online at https://doi.org/10.1007/s11187-019-00209-x

10.6.2019

Abstract

Although conventional register and survey data on entrepreneurship have enabled remarkable insights into the phenomenon, the added value has slowed down noticeably over the last decade. There is a need for fresh approaches utilising modern data sources such as Big Data. Until now, it has been quite unknown whether Big Data actually embodies valuable contributions for entrepreneurship research and where it can perform better or worse than conventional approaches. To contribute towards the exploration of Big Data in entrepreneurship research, we use a newly developed dataset based on publications of the German Press Agency (dpa) to explore the relationship between news coverage of entrepreneurship and regional entrepreneurial activity. Furthermore, we apply sentiment analysis to investigate the impact on sentiment of entrepreneurial press releases. Our results show mixed outcomes regarding the relationship between reporting of entrepreneurial events, i.e., media coverage, and entrepreneurial activity in German planning regions. At this stage, our empirical results reject the idea of a strong relationship between actual entrepreneurial activities in regions and the intensity of it being reported. However, the results also imply much potential of Big Data approaches for further research with more sophisticated methodology approaches. Our paper provides an entry point into Big Data usage in entrepreneurship research and we suggest a number of relevant research opportunities based on our results.

Keywords: *entrepreneurship, media coverage, mass media, Big Data, sentiment analysis, GEM, entrepreneurial ecosystem, region, news data*

1. Introduction

New, vast amounts of data and data sources for scientific research have become available in recent years and seem to be ripe for the taking, i.e., to be analysed with sophisticated algorithms and Big Data approaches. Big Data is a reality, and it needs to be harvested for scientific purposes. This includes entrepreneurship research. Yet so far, relatively few efforts have been made in this direction. Most of the research in entrepreneurship still relies on insight from traditional data sources like registers and surveys. Investigating value and possibilities of all kinds of new and promising Big Data sources to unveil novel insights into entrepreneurship, however, seems to be the necessary next step. Although in the long run both traditional survey data and Big Data might complement each other and coexist, Big Data might be the crucial new approach to moving our research forward now. This paper makes such a contribution by analysing the relationship between entrepreneurship and news coverage in public media. Thereby, we are not only presenting an application of Big Data, but new data as well.

Until now, it is quite unknown whether Big Data actually embodies valuable contributions for entrepreneurship research and whether it can perform better or worse than conventional approaches. This issue has gained importance for several reasons. Empirical- (or evidence)-based research on entrepreneurship increased in relevance within the domain of entrepreneurship research in recent decades (see, e.g., Audretsch 2012). Until now, the main source for quantitative data on entrepreneurship has been large-scale surveys or register-based approaches (see e.g., Coviello and Jones 2004). However, the former in particular requires significant investments of efforts and resources. Conducting statistically representative surveys, perhaps even with standardised questionnaires in different countries or sub-national regions and for a longer time period, is a challenging task that is not easy to fund and maintain. Consequently, searching for less expensive and easier methods to collect data on entrepreneurship is a major task to advance this field of research, especially in light of new methods, fast growing data sources, amounts and availability. But cost and accessibility of data are not the only relevant motivations for exploring new sources. Although entrepreneurship research has gained much from exploiting quantitative survey and register data, many new findings seem to be small increments building on the existing knowledge base. A good indicator for the saturation of a specific field of research is whether a new or rather newly packaged concept such as entrepreneurial ecosystems is introduced and blows up. This field offered more or less the illusion of something completely new. Publication statistics show that a huge amount of research effort was shifted towards this topic (Alvedalen and Boschma 2017). The measurement of entrepreneurship in its entirety needs to be revisited and modernised, as do its different components and influence factors. Due to digitalisation and internet-based platforms, Big Data allows for several new opportunities to create unique and specific databases in the near future.

This paper explores usage of news coverage for entrepreneurship research by examining the relationship of regional entrepreneurial activity with news and their sentiments using a Big Data set scraped from the web portal of the German Press Agency (dpa) subsidiary 'news aktuell'. The webpage has about 65,000 subscribers, mainly journalists and bloggers. Our aim is to explain the spatial pattern of entrepreneurship-related newsworthy events, based on more than 100,000 press releases scraped between May 2016 and November 2018 using access supplied by the dpa. The press releases have been explored regarding their statistical relationship with conventional indicators of entrepreneurial activities and media coverage. This particularly concerns information on entrepreneurship activities and the perception of entrepreneurship news coverage collected by the annual Adult Population Surveys (APS) as part of the Global Entrepreneurship Monitor (GEM). It leads to an interesting interpretation regarding the data quality of this particular, perception-based GEM variable. In addition, we analyse the entrepreneurship-related press releases with respect to regional differences in sentiments, i.e., if there are systematic variances in the way entrepreneurial activities are reported in press releases. Thus, our paper may be interpreted as a comparison of traditional, survey-based entrepreneurship data and Big Data.

Public media, even if reduced to news publications, show almost all of the related characteristics of Big Data. Following Kitichin and McArdle (2016), Big Data can show different sets and combinations of attributes or trait profiles. News data have a massive volume, even if only the number of articles is considered. But each article itself delivers a sub-level of additional information. Broken down into paragraphs, sentences, word combinations or just sheer word counts, the data multiply manifold. Filtering, matching or analysing this manually is impossible. Digital news data also have velocity. Different sites are competing for readers. News stories from yesterday are old and have lost their journalistic worth. Fast or even almost instant response time of news articles to real-world events has become the norm. Variety is a given as well. News data may be neutral reporting, of suggestive essayistic nature or ironically toned. They cover a huge variety of potential influences and topics aimed at different target groups and varying political spectra. Even if different sites report the same event, articles may differ quite strongly. As such, news data show all 3Vs described by Laney (2001) and Kitchin and McArdle (2016). Furthermore, they show signs of indexicality (because each article is unique and has a known source, time and date), relationality (it can be matched with other data sources as shown in this paper), scaleability (in close relationship to its velocity), veracity (in many cases, it is produced by humans who—especially in these large numbers—do not work flawlessly, and it is messy in terms of localisabilty, focus, quality and sourcing). News data are also valuable because they contain many layers of information (Mayer-Schonberger and Cukier 2013, Dodge and Kitchin 2015, Marz and Warren 2012, Marr 2014). For an ontological overview of these terms and a comparison between 'small' and Big Data types, see Kitichin and McArdle 2016.

To test the news data for entrepreneurship research usability, we explore the complex relationship between the factors influencing media attention and entrepreneurship in sub-national German regions on the one hand and the measurement of entrepreneurship activities on the other. Measurement of entrepreneurship activity in a region may use a wide variety of conventional or innovative data, both direct or indirect, either from survey or register-based sources. We argue that the potential of media data is large and almost completely unexploited. It delivers not only count data, but also text bodies with vast opportunities for research. Sentiment analysis is another new method we apply in our paper. Does positive/negative reporting influence entrepreneurial activities? Does this even differ systematically across space? If so, is it based on cultural differences venturing into path dependence and context, or does the difference stem from individual actors distributed by chance? This paper cannot answer this multitude of questions, but rather highlights an entry point by showing ways of exploring the usefulness of new Big Data sources for entrepreneurship research with extension to the spatial dimension. The focus on the spatial level of sub-national regions is used because it is the most relevant geographical context dimension for entrepreneurial activities, as numerous scholars have shown (e.g., Sternberg 2009; Feldman 2001).

Our research is explorative in terms of the core indicators used because this is the first attempt to apply the German web portal 'Presseportal' of the dpa (German Press Agency) subsidiary 'news aktuell' for research on regional entrepreneurial activity and media coverage. Since research on this topic is still rare (see also Wang et al. 2017), opening up new approaches through Big Data could lead to much-needed progress.

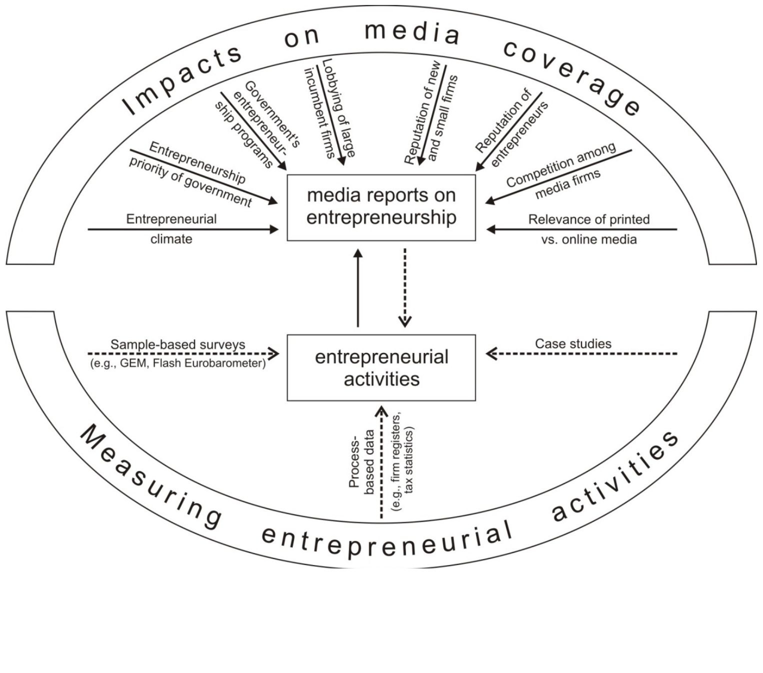
The remainder of the paper is structured as follows. We start with an overview of why media coverage and regional entrepreneurial activity might be related. Section 3 provides a description of the data and methodology used. Our empirical results together with a discussion are presented in Section 4. Section 5 concludes.

2. Entrepreneurship and media—some conceptual thoughts and the role of Big Data

The relationship between media and entrepreneurial activities, both seen from a regional perspective, is rather complex and interdependent (see figure 1). Two main directions may be distinguished. First, media coverage is, besides other determinants, influenced by real entrepreneurship activities in the territory, although not all entrepreneurship activities occurring in a given region will be considered newsworthy by public media, and regional media will not report about regional entrepreneurial activities exclusively.

Second, the extent and kind of entrepreneurship activities and entrepreneurship attitudes in a given region are influenced by the context factor of 'media coverage' because media are noticed by real and potential entrepreneurs and the latter's entrepreneurial behaviour depends on the individual's perception of media news about entrepreneurship. Of course, media coverage is but one of those many factors influencing and being influenced by entrepreneurial activity, but it has hitherto rarely been used even as an outcome or proxy variable of entrepreneurship (e.g., Amodou et al. 2016).

Fig. 1: The relationship between media coverage and entrepreneurial activities


Impacts on media coverage
Government's entrepreneurship programs
Lobbying of large incumbent firms
Reputation of new and small firms
Reputation of entrepreneurs
Entrepreneurship priority of government
Competition among media firms
media reports on entrepreneurship
Entrepreneurial climate
Relevance of printed vs. online media
Sample-based surveys (e.g., GEM, Flash Eurobarometer)
entrepreneurial activities
Case studies
Process-based data (e.g., firm registers, tax statistics)
Measuring entrepreneurial activities

Thus, public media may be considered a context factor that, besides many other context factors and person-related factors, influences entrepreneurial activities as well as entrepreneurial attitudes of individuals or, at the aggregated level, of regional or national economies or societies. In that sense, media coverage of entrepreneurship can be considered an informal institution of entrepreneurial regions (Glaeser et al. 2016; Obschonka 2017). Analysing it as interdependently related to entrepreneurial activities is not necessarily new, but if it is considered as input, context and result of entrepreneurial activity, it is worth a closer look due its potential as a motoric unit to push endogenous growth through new firm formation. While we do not investigate the role of media coverage as a context factor for entrepreneurial decisions, activities or attitudes in this paper, the new indicators discussed here are promising candidates to be used in future research exploring the relationship between said dimensions of entrepreneurship and this rather rarely used factor in empirical entrepreneurship research.

Related to this perspective on the relationship between media coverage and entrepreneurship activities, we introduce Denzau and North's (1993) sender-receiver model as a second theoretical foundation of our main argument, and it can be combined with the (regional) context argument explained above. The sender-receiver model has recently been introduced to regional entrepreneurship research, related to regional role model effects in particular (see Wywrich et al. 2016, 2018). For our purposes, the mainly non-social interactions between entrepreneurs or entrepreneurship activities in a given region on the one hand and, on the other hand, those who write about such activities (journalists, bloggers and the like) can be explained with the help of a sender-receiver model. The entrepreneurs, their activities and the related events are the senders that transfer signals to the receiver, who is an observing journalist or blogger. Depending on the personal perception, the signals are interpreted in a positive or in a negative way, but rarely as neutral. The way he/she interprets them also depends on the context he/she is living and working in (e.g., previous experiences with entrepreneurship or reactions to his/her previous publications regarding entrepreneurship). In a next step, of course, the news the journalist writes may have effects on real or potential entrepreneurs who read it. The news will be supportive or a hindrance with regard to entrepreneurial intentions or activities. However, these latter processes are not the main focus of our paper, but will be addressed in the named sentiment analysis. The application of this sender-receiver model to news producers provides fruitful connections to the entrepreneurship literature on (regional) opportunity recognition (see, e.g., Arenius and Minniti 2005, Stuetzer et al. 2014), applied to news producers instead of entrepreneurs. Note that the ambivalence of these signals should be considered: the same signal (e.g., a given entrepreneurial action) may be perceived very differently by different receivers, so the medial consequences might be very different, too, depending on the perception of the journalist, the blogger or others who produce the news (or decide not to write a story about it). Some of these processes are, at least partially, psychologically driven and are not just person-specific, but also region-specific (contexts are space dependent), as recent empirical research by psychologists with respect to regional entrepreneurship has shown (e.g., Fritsch et al. 2018 on German regions or Ebert et al. 2018 on the correlation between courage and entrepreneurship in US regions). In our paper, the opposite direction matters most: how are entrepreneurship activities covered by the media, namely news media? In other words, is entrepreneurship news a good indicator of real entrepreneurship activities at the regional level? Because both directions of these effects are interdependent, they have both been included in figure 1.

Measuring the impact of media on economic events and vice versa, massively harvesting and analysing public media coverage of those events are still surprisingly unused in academia to our knowledge. Various studies have developed evidence for older states of mass media by which media coverage of news on economic processes or events impacts the very nature of the subject itself (Coyne and Leeson 2004; Goidel and Langley 1995; Doms and Morin 2004; De Boef and Kellstedt 2004; Wartick 1992; Carroll and McCombs 2003). We argue that on the one hand, today's media show certain similarities to this, but on the other hand, they have a completely different dimension in quantity and quality. This is the case for both digital media that is, in regards to the state of research, new and analogue news media such as daily newspapers. Media need to be revisited due to the ubiquity of available access to news, posts, tweets and 'stories' of vastly different frequency, content, tone and quality. Furthermore, regional aspects have to be explored but are not as easily attributable as classic newspapers. A rewarding field of study lies ahead.

More recent impact of news or media coverage with links to Big Data can be drawn from the narratives literature. Shiller (2017, p.49) states that '*research [...] needs improvement in tracking and quantifying narratives'*, which we attempt in this paper. Narratives may help us to understand the relationships of news reporting and regional entrepreneurial activity as Roundy (2016) showed, theorising entrepreneurial ecosystems (EES) as both sources of narratives and being influenced by them, e.g., through role model display (see also Spigel 2015). Research in this new field of the discipline, the systemic view of interdependencies of entrepreneurial activity and (regional) context factors (and between the latter themselves), is based predominantly on traditional data sources. However, with the slowdown of breakthrough discoveries, it is necessary to look for ways to picture entrepreneurship from other sides by exploring new data sources.

When it comes to EES, there is a significant research gap in many aspects of the phenomenon due to missing reliable (quantitative) empirical longitudinal and cross-sectional data. The role of media or narratives in EES is no exception. As an important medium to broadcast success stories, to push constant visibility of entrepreneurial related events or to boost a region's spirit and culture towards a 'start-up mentality', it could potentially play a crucial role. However, empirical evidence to assess the actual causal impact on, e.g., the total amount of start-ups in a given region is scarce and would probably require more qualitative than quantitative methods and data.

Regional (sub-national) approaches to analysing media coverage come with additional challenges. The quality, frequency and focus of reporting and its impact and the digital availability of regional news sources and their localizability can vary considerably. However, for academic disciplines such as economic geography or regional science, this opens a promising new line of questioning to identify spatial patterns and causal relationships between regional economic processes or events and news coverage of those events in both directions. Mass media coverage in the form of news could potentially be used to estimate different kinds of space-sensitive context factors leading to regional entrepreneurial activity and the activity itself. Not only context factors, but also different kinds of entrepreneurial activity interact with media coverage. Regional differences in media coverage of entrepreneurial events may be rooted in a number of possibilities: a high impact, innovative or fast scaling start-up will probably receive more media attention than necessity-based new firm formation without an innovative idea, a small budget and limited human resources. In general, everything deviating from day-to-day news might be considered newsworthy.

Other conditions resulting in spatial differences of news coverage could be the overall level of entrepreneurial activity, the economic sectors and their shares, different stages of general economic development or simply timing-based conditions such as extraordinary public events or shocks. Additionally, sub-national regions with high density of newsworthy events may undercover, e.g., success stories of new start-ups, which would be headline news in regions with a low density of newsworthy events. Those, in turn, might cover such events more prominently than what would seem proportional. However, certain aspects could lead to masking effects of media coverage or even completely negate them. One advantage for Big Data approaches to news, the amount of produced data itself, could very well have a negative downside by leading to overstimulation of individuals and thereby numbing of reception and ultimately reducing the impact. Additionally, the need of mass media suppliers to produce high impact, sensational articles at high speed and frequency may lead to overestimating events. For older states of mass media, it could be shown that the news stories did not always cover the economic realities (Blood and Phillips 1995; Goidel and Langley 1995; Fogarty 2005). This has to be kept in mind when dealing with current media output as well. Nevertheless, there is a relationship between news coverage and entrepreneurial activity (Hindle and Klyver 2007). As pointed out, the relationship is interdependent. However, in this paper, we deliberately focus the empirical part on just one side of this causal relationship by exploring the degree of presence of entrepreneurship-specific news impacted by entrepreneurial activity in regions. The other causal direction, the influence of media coverage as a context factor on entrepreneurial activity (e.g., via role model visibility), has received at least some attention (Greenwood and Gopal 2017; Hindle and Klyver 2007). Recently, and predominantly in discussions in which media are important pillars in entrepreneurial ecosystems (EES) (see e.g., Isenberg 2010), influence of entrepreneurial activity on media coverage is vastly under-researched to our knowledge (Hang and Weezel 2007). One of the few studies on this side of the narrative, albeit from a different angle, is by Amodu et al. (2016). To explore the nature of news coverage on entrepreneurship, they collected articles from four newspapers over the course of three years and analysed them for news on entrepreneurship using content analysis. However, the findings remain descriptive at best, arrived at by categorising and counting the topics of the identified articles. The findings were not set into relation to actual entrepreneurial activity and were not classifiable as a Big Data approach.

Hindle and Klyver (2007) looked into the opposite side of the causal relationship between media coverage and entrepreneurial activity. By relating GEM data for entrepreneurial activity and motive (opportunity and necessity) and perception of news coverage, they found a weak but significant impact of perceived media coverage of entrepreneurship on opportunity-driven early-stage entrepreneurial activity and on owners of young businesses. However, they urge interpreting these results with care. They refer to the reinforcement model from media theory, arguing that mass media are only capable of reinforcing their audiences' existing values and choice propensities, but are not capable of shaping or changing those values and choices, i.e., media news would be unable to increase or decrease entrepreneurial activities in a given region. These authors, however, use a variable that covers the perception of media news impact on entrepreneurship, not the media news themselves. The idea of 'changing' versus 'reinforcement' versus 'shaping' is based upon their extensive review of mass communication theory literature. Organising this literature into these categories, they display three partly contradictory hypotheses on media effects on entrepreneurial behaviour that each had a dominant period in history.

Greenwood and Gopal (2017) found that temporarily higher coverage of specific news may lead to an increase of entrepreneurial activity related to this particular field. We argue that in the long run, our data can help isolate and analyse specific singular aspects of entrepreneurial activity as soon as the data base has increased to sufficient size.

To sum up the state of research in terms of the relationship between media coverage (and its perception) and entrepreneurship activities, empirical research is rare in general, and no Big Data attempts are known to the best of our knowledge. If empirical evidence is available, it focuses on the effects of media coverage on entrepreneurship, but not on the opposite one. Thus, this is the focus of our paper.

In light of recent developments in the fields of machine learning, social media, information storage and availability of huge amounts of untapped data, the way to collect data for scientific research on entrepreneurship needs to be reviewed and challenged (see also Mahmoodi et al. 2017). New approaches and data sources may be worth investigating separately from and in addition to conventional ones, which enables research to uncover hidden aspects that could not be captured with standard survey designs before. Mahmoodi et al. (2017, 58) state, *'An integration of Big Data and traditional approaches might help to optimize both the prediction and explanation of behavioural phenomena'*. This is underscored by Big Data application in recent (social) science studies, such as Obschonka (2017), Kosinski et al. (2016), Chen et al. (2017), Wang et al. (2017) and Glaeser et al. (2016), to name but a few.

Coviello and Jones (2004, 485) in their overview of '*methodological issues in international entrepreneurship research*', found that the majority of data gathering approaches in this field are quantitative surveys. Such survey-based approaches can yield reliable, high quality comparable data for many countries (e.g., GEM), but nearly all of them have at least some shortcomings. Surveys come at high costs and need manpower to be completed. They cover perceptions of respondents, not facts, and are therefore weak in subjectivity. Transparency and reliability are often difficult to achieve, especially when representative surveys are conducted to collect data on rare events like entrepreneurial activity within the population of regions. If not repeated with necessary frequency, surveys cannot cover dynamic processes or different stages. Furthermore, it takes a lot of time to build a questionnaire and conduct the survey. These complications may lead to sample sizes that are smaller than optimal, a point often argued when it comes to claiming representativeness.

As with perception data, media or news data probably do not depict reality but rather an interpretation of it, which might be glorifying, suggestive, apologetic, hostile or any other form of subjective picture. Depending on the source of the news, there might be a hidden agenda. Setting different news sources into relationships with their content and the sentiment in which they are displayed could shed some interesting light on processes otherwise only discoverable with qualitative in-depth case studies, but for quantitative data. To dive deep in the natural text processing and analysing of the individual articles, an even more sophisticated dataset containing different sources for news, especially regional coverage, is necessary. To produce and explore this will be the next step in our research, built upon our current findings. But as a first step, we apply a sentiment analysis for the positive and negative dimensions because news stories can be good or bad, but they are seldom neutral (Dodbole et al. 2007). The tone of reporting matters for how individuals perceive specific events. For instance, the frequency of negative news lowers consumer confidence below what economic fundamentals would suggest (Doms and Morin 2004; Hollanders and Vliegenthart 2011a). Changes in corporate reputation similarly are explained by media exposure (Wartick 1992; Fombrun and Shanley 1990; Carroll & McCombs 2003). Hence, besides their (information) content, the frequency and tone of news coverage also influence agents' economic decision-making, of which entrepreneurship is one.

Entrepreneurship represents interesting events that may serve as inputs for journalistic production. The relevance of this input may significantly vary among regions based on many different factors. Most importantly, we expect the frequency of the entrepreneurial event to be a factor in this context. In regions in which entrepreneurship is rather uncommon, such events may receive higher journalistic attention. We do not argue that entrepreneurship in itself is necessarily sensational, but rather that its (in)frequency may produce this characteristic. Journalists in regions with high levels might be less prone to cover each new start-up or idea simply due to them being common. They may not possess the characteristics of a 'sensation' and hence receive comparatively less attention (see also the call of Welter et al. (2017) for more academic attention to 'everyday entrepreneurship' that would surely not be covered by the named types of media). This effect may be counterbalanced by the generally higher frequency of entrepreneurship events. While in these cases a smaller share of events may find its way into the news, the larger share of events may still lead to a higher (absolute) coverage. There is much uncertainty in this, which uncovers a dire need for more research.

As a first exploration, we focus in our paper on a unidirectional impact of entrepreneurship activity on news reporting. This does not cover the complete picture of the causality between these variables, but serves as a venture point to dive deeper into the complex linkages.

3. Empirical approach

3.1. Dependent variables: newsworthy entrepreneurial events and sentiments of reports on entrepreneurial activities

We rely on data collected by the German website www.presseportal.de. Presseportal is the web portal of the dpa subsidiary *news aktuell*. It is the largest and most popular PR portal in Germany, with about 9 million visitors per month and over 12,000 companies being represented with their own newsrooms. The webpage has about 65,000 subscribers, mainly journalists and bloggers (Presseportal 2018). Accordingly, our data do not represent news appearing in newspapers or social media, but rather information that actors want to share and would like to see being picked up by a wider audience and that they seek to be distributed by different kinds of influencers and news distributors.

Unfortunately, we do not know which press releases or which share thereof are actually published in newspapers or on social media platforms. This has significant implications. Most importantly, the data are not representative for the actual news coverage in regions or with respect to specific topics. We do not even know to what extent they correlate to what readers might find on average in newspapers or other news outlets. However, a press release is one of the most important PR tools and provides journalists with their raw material, which is regular, reliable and usable information (Walters & Walters, 1992). Our dataset therefore provides a detailed picture of what newsworthy events take place in a region. Notably, this picture is taken before professional journalistic editing and selection. In this case, newsworthiness is determined by actors responsible for or participating in events, which implies that news is, ultimately, 'not what happens, but what someone says has happened' (Sigal, 1986). In our case, the 'someone' is not the journalist, but actors issuing press releases. In summary, the data contain information on events for which actors believe a certain public interest exists and that have a chance of being picked up by different sorts of news outlets. This has to be kept in mind when interpreting the results.

We downloaded the news data from the webpage for somewhat more than two years (May 2016 to November 2018). While at first we were able to get data entries that were a few months old, in early 2017, the access was restricted by dpa to downloading a maximum of 1,000 releases from the point of time of scraping. We restructured the data gathering to be done on a daily basis. The press releases were automatically accessed, downloaded and processed into a database.

In total, we retrieved 100,701 press releases, which corresponds to about 2,800 releases per month. The releases contain a unique id, a title, a fixed URL, a text body, the date of publication, classification into one of six broader topics ('financial', 'economics', 'politics', 'sport', 'culture' and 'miscellaneous'), a list of keywords and an identification number for the publishing actor. Unfortunately, the keywords and broader topics proved to be of rather general nature and hence of little value. We therefore focus on the text body to obtain the information of interest: location and content.

To extract locational information from the text, we first obtained a list of all places (settlements, villages, towns, cities) in Germany from the OpenGeo-database (http://opengeodb.giswiki.org/wiki/OpenGeoDB). The database contains the names and geographical coordinates of more than 11,000 places in Germany.

The geographical geolocating was a multistep procedure. First, we extracted the information about the location of the press releases' newsrooms, i.e., the location of the actor that submits the press release to the Presseportal. This is consistently given at the beginning of the text. However, this information does not necessarily refer to the exact location where the event the press release is informing about actually took place. We therefore extracted additional locational information from the remaining body of text. Using a string matching procedure, we identified all words potentially indicating locations in the text. Subsequently, the potential locations were checked and in some cases adapted to allow for unique matching. This particularly applies to city names that are combinations of multiple words such as the city of 'Frankfurt am Main'. Here, alternative versions of the location's name exist, e.g., 'Frankfurt/Main', 'Frankfurt a. M.', 'Frankfurt a. Main', which had to be identified and harmonised. Another problem is names that refer to multiple (distinct) locations. For instance, the name 'Halle' may refer to the city 'Halle an der Saale' or 'Halle (Westfalen)'. In these cases, we searched the texts for additional information giving indications on the correct location, for example, the name of the federal state or surrounding region (e.g., 'Sachsen-Anhalt'). More problematic cases are location names with multiple meanings, with the city of 'Essen' being a prime example. It literally translates into 'food' or 'eating', a word that appears at relatively high frequencies in the releases without referring to the city. Lacking an alternative, we had to drop such cases, which means that certain locations (foremost 'Essen') do not appear in our empirical analysis. In future research, more advanced matching algorithms might be able to deal with these cases. Our approach also implies that multiple locations can be assigned to a single article. In fact, this happens relatively frequently with press releases referring on average to almost two (1.9) locations. The reason for this is that press releases frequently apply to multiple locations or they deal with relationships between locations. Moreover, some releases cover multiple events in different locations. To ensure a high quality of the location procedure, we split the press releases into paragraphs. Press releases have on average about 4.9 paragraphs. If a location was found in a paragraph, it will be assigned to this paragraph (and the content therein). If no locational information was found, the location of the release issuer (the organisation behind the corresponding news rooms) was assigned to the paragraph. We identified (entrepreneurial) events in the press releases at the level of paragraphs as well. Accordingly, the link between events and locations are established at this level. Nevertheless, when counting events, we count at the level of full press releases to avoid discriminating single versus multiple paragraph releases. In practice, this means that a press releases will be counted multiple times when the information on a specific (entrepreneurial) event and location coincide in multiple paragraphs or when multiple locations or events are found within the same paragraph. However, this occurs only in 3% of all cases. While there are just 1.2 locations per paragraph in general, we identified about 1.5 locations on average in paragraphs containing entrepreneurial events.

In total, we have been able to assign at least one location to 85,439 press releases, which corresponds to a success rate of almost 82.6%. Based on the location information, we aggregated the press release information to the 96 German planning regions defined by the BBSR. Figure 2 shows the geographical distribution of press releases in those regions. Clearly, most press releases are reported in cities, with the absolute population being a very good predictor of the number of press releases (r=0.82***).

To identify press releases relating to entrepreneurial events, we compiled a list of 69 words clearly referring to entrepreneurial activities. We filtered the press articles according to these words. In total, 2,952 press releases and 5,887 paragraphs included at least one of these words. Table 1 shows the ten most frequent keywords. Note that multiple keywords may be found in one press release.

Fig. 2: Map of spatial distribution of press releases

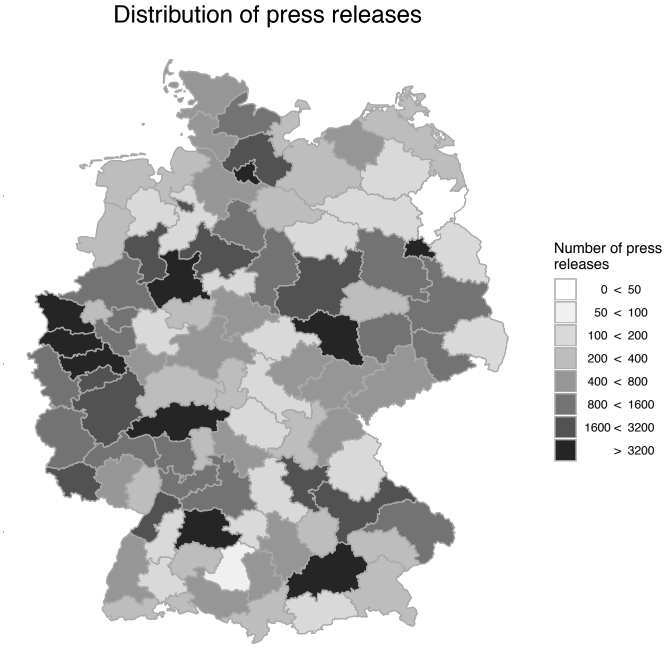


Table 1: Keyword frequencies

RANK	KEYWORD	COUNT
1	Startup & start up	2160
2	entrepreneur	250
3	unternehmertum	186
4	unternehmensgründ	174
5	accelerator	150
6	venture.capital	125
7	Junge unternehm & junguntnernehm	217
8	existenzgründ	96
9	inkubator	82
10	risikokapital	81

Figure 3 visualises the spatial distribution of entrepreneurial press releases by colouring regions according to the share of releases referring to entrepreneurial events. Interestingly, the strong relationship to population disappears, and rather rural regions in proximity to urban regions seem to be characterised by large shares. However, the distribution is rather inconclusive and demands more comprehensive analyses.

We use the press-release information to construct our dependent variables. The first one is ENTRE_COUNT, which is the number of press releases that include at least one of the keywords related to entrepreneurial activities. It captures the frequency with which entrepreneurial events are considered newsworthy and are consequently featured in press releases. Alternatively, it can be interpreted as the intensity of regional entrepreneurship events being fed into the journalistic process and hence, potentially, being covered in news.

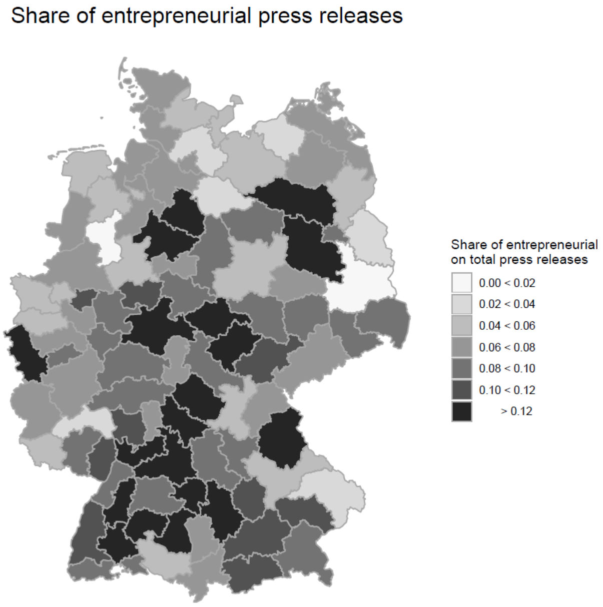
Our second dependent variable captures the way entrepreneurial events are reported in press releases. Sentiment classification is the task of determining the overall sentiment orientation (if any) in a text (Ohana and Tierney 2009). In this study, we apply a lexicon-based sentiment polarity categorisation approach. That is, we consider the text as a collection of its words, disregarding the grammar and word order (a so-called bag-of-words approach). Among these words, we count the number of words associated with 'positive' and 'negative' sentiments in each press release. Crucially, we need a definition of words' polarity, i.e., their degree of 'positiveness and negativeness'. In a common manner, we rely on a list of words that are pre-coded with respect to their polarity (Taboada, Brooke, Tofiloski, Voll, & Stede, in press). For this paper, we utilise the SentimentWortschatz (SentiWS), which is a publicly available German-language opinion lexicon listing positive and negative polarity-bearing words, with polarity values ranging between [-1,1] (Remus et al. 2010). To apply this approach, we first clean the texts and remove unwanted characters, addresses, links, numbers, punctuation and German stop-words. Second, a document term matrix is created and weighted with the polarisation values of the sentiment dictionary (*sentiWS*) using a linear model (Feuerriegel and Proellochs, 2018). That is, each press release's sentiment score is the sum of polarization weights of the word tokens. Their aggregate at the regional level (SENTIMENTS) represents the average sentiment of a region's press releases, which will serve as a control variable in some empirical models. By restricting the sample to press releases containing entrepreneurship-related keywords, we create the second dependent variable (ENTRE_SENTIMENTS) giving insights into regional variations in sentiments of entrepreneurship press releases.

While this is a relatively simple approach, it is efficient and capable of achieving high levels of accuracy (Richard and Gall 2017). Given that we apply this approach to more than 85,000 texts totalling more than 74,067,694 words, efficiency in particular is a crucial dimension in this context. Of course, efficiency comes at some costs. For instance, we cannot handle negation and intensification. Moreover, we concentrate on polarity and do not consider other sentiments such as anger and dislike. Crucially, we also do not know the extent to which the sentiments are directly associated to the entrepreneurship events reported in the press releases. If these are not the primary content of the releases, our analysis is likely to associate sentiments concerning other topics to these events. Accordingly, our measure reflects the general sentiments in press releases that also refer to entrepreneurship events. This has to be taken into consideration in the interpretation of our results and also indicates potentials for improving this type of analysis in future studies.

3.2. Explanatory variables

To explore the relationship between news and regional entrepreneurial activities, we rely on GEM data. More precisely, we use an excerpt of a unique set of regional GEM data for Germany. By pooling annual representative national Adult Population Survey (APS) data (respondents were between 18 and 64 years old), including location information of respondents, we circumvented the insufficient number of cases per year and region that usually troubles sub-national level analysis with these data. The result is a unique regional GEM dataset for Germany. A standard GEM APS consists of at least 2,000 cases per country and year.

Fig. 3: Map of share of entrepreneurial news


Share of entrepreneurial press releases
Share of entrepreneurial on total press releases
0.00 < 0.02
0.02 < 0.04
0.04 < 0.06
0.06 < 0.08
0.08 < 0.10
0.10 < 0.12
> 0.12

For this paper, we used a dataset spanning the years 2012-2017. Although we have an implicit time lag to our more recent press release data, we argue that this poses no limitations on our analysis due to the nature of entrepreneurial activity in Germany. The consistently low level of entrepreneurial activity (compared to other innovation-driven countries) has been rather stable over the last decade. Although entrepreneurial activities vary somewhat over time, these variations are rather negligible in terms of magnitude across years. In fact, in most cases, the annual differences are not statistically significant, thereby allowing the pooling of annual data into a cross-sectional dataset (see Sternberg et al. 2018). It also implies that we do not need to consider a time lag between press release information and entrepreneurial activities and assume as well treating both as time-invariant. Future research should nevertheless more systematically explore potential long-time variations.

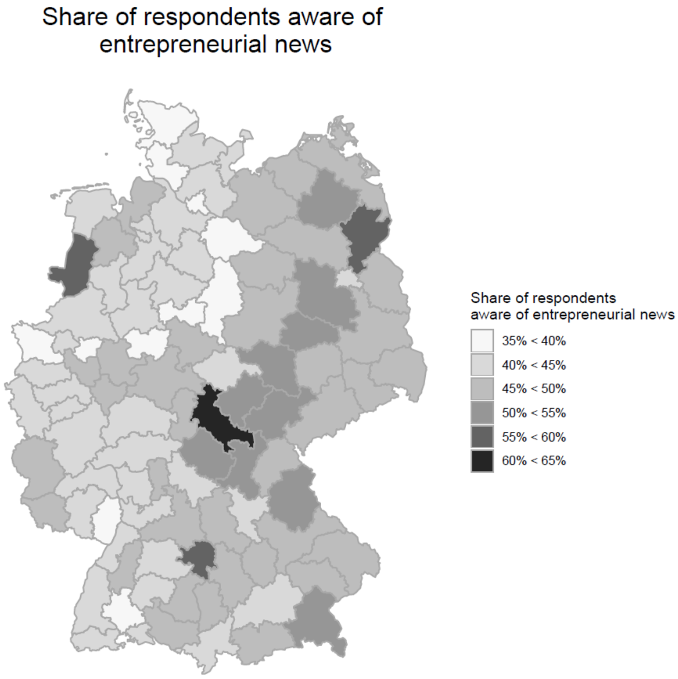
We use the GEM data to create two variables. The first is FOUND, which is based on the quota of the GEM variable TEA (Total Early-stage Entrepreneurial Activity). It is calculated by the number of respondents with TEA='yes' divided by the number of respondents with TEA='no' for each region. A respondent is considered in the estimation of TEA when he/she is either a nascent entrepreneur actively pursuing a business foundation or if he/she manages a business less than 42 months old. Accordingly, the measure captures recently founded businesses and start-up intentions. It represents an excellent proxy of entrepreneurial activities that takes into account that entrepreneurship is a process and not a status. We therefore argue that TEA measurement of GEM is superior to alternative approaches, foremost those using registers that exclusively cover successful entrepreneurs.

The GEM data offer a range of additional information on demographics and entrepreneurship intentions. In the context of this study, we are particularly interested in information on the perception of (entrepreneurship) media coverage that is also included. More precisely, we focus on the variable NBMEDIA. It summarises the respondents' answers to the statement '*In [your country], you will often see stories in the public media and/or internet about successful new businesses*'. The answer can either be yes or no. The GEM variable is worded towards the national scale, but we argue that the respondents' actual perceptions are influenced by their regional context. Although this allows us to use this variable for our analysis, it may lead to a reliability problem of NBMEDIA for national level calculations. By including it in this analysis, we try to estimate the quality of NBMEDIA as proxy for covering news on entrepreneurship alongside our main focus. We create a regional aggregate (MEDIA) indicating the share of positive answers on total regional respondents. Figure 4 illustrates the spatial distribution of this variable. It indicates the existence of an East-West discrepancy, with values in East Germany (former GDR) (av. 49.98) being on average higher than West-German regions (av. 44.3)[1]. However, this may in part also be explained by the East's lower degree of urbanisation because this measure correlates negatively with population density at r=-0.27***.

Our first major control variable is based on press release information and counts the number of press releases per region (RELEASES). It primarily serves as a control for variations in regions' general likelihood to appear in press releases. It captures potential differences in the occurrence of newsworthy events and variations in the propensity that these events will translate into press releases.

[1] The difference is statistically significant at 0.01, as indicated by a two-sample t-test.

Fig. 4: Map of share of respondents aware of entrepreneurial news


Share of respondents aware of entrepreneurial news
Share of respondents aware of entrepreneurial news
35% < 40%
40% < 45%
45% < 50%
50% < 55%
55% < 60%
60% < 65%

In addition to press release information and GEM, we consider a list of control variables that account for fundamental differences between regions that might impact the relationship between entrepreneurship and press releases. These data are taken from official German statistics accessible via www.inkar.de. This database collects information from national and federal statistical offices in Germany and supplies numerous indicators for various spatial scales. Population size (POP) and population density (POPDEN) control for absolute numbers as well as agglomeration and urbanisation factors, which can strongly influence entrepreneurial activity (see Bosma and Sternberg 2014). GDP per capita (GDP_PC) controls for economic strength, and STUDS equals the amount of students at higher-education institutions per 1,000 inhabitants, being a proxy of knowledge creation and potential university existence and spin-offs. UNEMPL are unemployed persons of all employable inhabitants. To take into account still existing fundamental differences between regions belonging to the former GDR (East Germany) and West-Germany, we include a dummy (EAST) that is 1 if a region is located in East Germany and 0 otherwise.

3.3. Methodology

The aim of the present paper is the analysis of the potential relationship of news and entrepreneurial activities at the regional level. That is, we want to explore to what extent variations in the levels of regional entrepreneurial activities are mirrored in news (as approximated by press releases). As pointed out, our first dependent variable is the number of press releases per planning region, which is a count variable. Figure 5 shows the variable's distribution, which clearly signals over-dispersion. This is confirmed by the over-dispersion test of Cameron and Trivedi (1990)[2]. Therefore, we employ a negative binominal regression[3]. We do not find any indications of spatial autocorrelation, multicollinearity and outliers. Given that few keywords dominate the identification of entrepreneurial news or press releases (Table 1: Keywords), we test the robustness of the results with respect to their selection. For this, we re-run the analyses based on press releases identified to contain entrepreneurship-related context based on the two most important keywords ('start-ups' and 'entrepreneur').

Our second analysis focuses on the explanation of regional variations in the sentiments characterising entrepreneurship-related press releases. The variable ENTRE_SENTIMENTS representing the average sentiments of such press releases in a region is continuous, non-truncated or censored. Accordingly, simple OLS regression is appropriate. When log-transforming the dependent and all independent variables, the according models meet most assumptions. There are no signs of multicollinearity and spatial autocorrelation. The normal distribution of the errors can also not be rejected (when considering all control variables). However, we have to exclude Berlin, which distorts the estimations as an outlier. Moreover, despite taking the log of all variables, heteroskedasticity cannot be rejected. We therefore employ robust standard errors. To further substantiate the estimations, we calculate the 95% confidence interval of the coefficients using a bootstrap approach with 1,000 replications. Lastly, we use a binary logistic regression on a binarised version of the dependent variable that is 1 if the mean sentiments on entrepreneurship-related press releases are larger (more positive) than the mean across regions and 0 otherwise.

4. Results and discussion

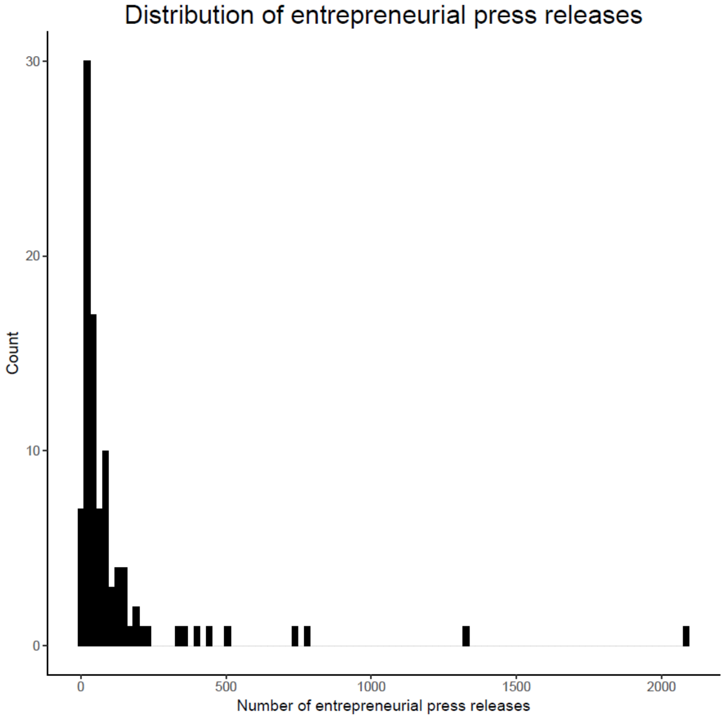
Table 2 presents the results of the negative binominal and Poisson regression analyses with the number of entrepreneurial press releases in a region as the dependent variable. Our baseline model is reported in the fourth (negative binomial) and seventh (Poisson) columns labelled 'general'. The previous three columns contain models of robustness checks, including different sets of explanatory variables. The fifth and sixth columns, labelled 'entrepreneur' and 'start-up', give insights into models in which the identification of entrepreneurial press releases is based on the two most common keywords.

With the different models, we try to understand the regional dimension of newsworthy entrepreneurship events with a particular focus on alternative measures of entrepreneurial activities. That is, we seek to explore the potential of news data as a new source for insight into regional entrepreneurial processes and activities.

[2] The dispersion statistic is d= 29.65 and z = 3.10 with the p-value = 0.0009732.

[3] For completeness, we also show the results of Poisson regressions. While the two models' coefficients are almost identical, the Poisson regression yields smaller standard errors and hence more statistically significant results.

Fig. 5: Distribution of entrepreneurial news per region



All models are reliable, and the Poisson estimations in particular show high pseudo-R2 values, suggesting that we are able to explain significant portions of the interregional variations in entrepreneurship-related press releases. Moreover, the coefficients of the control variables correspond to our expectations by and large. The most important predictor of entrepreneurship-related press releases is the total number of press releases in a region (RELEASES). Accordingly, the more events take place in a region (or are reported about), the more frequently entrepreneurial events are among them. Population (POP) relates positively to the number of entrepreneurship-related press releases. However, population density obtains a significantly negative coefficient, suggesting that in urban regions, either entrepreneurial events are less likely deemed newsworthy or that less of these events take place. Given that the latter contrasts with a well-accepted fact in the literature (e.g., Bosma and Sternberg 2014), entrepreneurial events seem to be perceived as less relevant to report about in urban regions.

We observe higher reporting rates of entrepreneurial events in regions with more higher-education possibilities (STUD) that are approximated by the share of (higher-education) students in regions’ total populations. Hence, in such regions, entrepreneurial events are more likely to be considered newsworthy and included in press releases. Most likely, this is due to higher frequencies of occurrences. The same applies to regions with higher levels of GDP per capita. This is likely related to opportunity entrepreneurship, which dominates necessity entrepreneurship in Germany.

Another interesting finding is that in regions belonging to the federal states of the former GDR (EAST), the share of entrepreneurial press releases is consistently higher across all models in which entrepreneurial press releases are identified with the occurrence of the word 'entrepreneur' alone. Anecdotal evidence suggests that the use of the word 'entrepreneur' is less common in this part of Germany. Accordingly, this finding may hint at cultural/historic differences in regional languages that are inherent to this type of data. A comparison of other text may substantiate this in a more systematic manner in future research. The sentiment analysis indicates that entrepreneurship news in East Germany have the tendency to show negative sentiments (Table 3). This shows that our data are able to display at least two different dimensions of impact, quantitative and qualitative ones.

The coefficient of our focal variable, FOUND, remains insignificant in negative binomial regression (except for the 'start-up' model). It becomes significantly negative in the Poisson analysis (also because of its negative link with the appearance of 'start-up' references in press releases). This finding implies that the reporting intensity of entrepreneurial events is not higher in regions in which more entrepreneurial activities take place. This contrasts with our expectations, but might be explained by a kind of customisation effect. We suspect that in regions in which entrepreneurship is a relatively frequent event, people perceive them as less newsworthy because they are rather common. Accordingly, they are relatively less likely to be found in press releases. If this effect also translates into lower news coverage of entrepreneurial events (our analysis only covers the input into the news process, not the actual output), this finding may have significant implications for innovative, start-up producing ecosystems that are typically located within larger cities and dense areas. Another possible explanation might be in line with the negative impact of population density (especially within the 'start-up' model). This effect is based upon urban areas showing higher numbers in absolute amount of non-entrepreneurial newsworthy events in relationship to entrepreneurial ones, which reduces the share of entrepreneurial news. Newsworthy entrepreneurial events may find themselves in stronger competition with a broader spectrum of other reportable events, such as political, sportive, etc., that are most likely overrepresented in such regions.

Our second focal variable is MEDIA. It gains a significantly negative coefficient in all models. We therefore find empirical support for less frequent reporting on entrepreneurial events in regions in which individuals indicate higher exposure to entrepreneurial news. Put differently, while subjectively individuals perceive relatively little news coverage of entrepreneurial activities in their region, we actually observe a comparatively higher reporting of such. This finding holds even when excluding all control variables and running the models in different specifications. The finding comes as a surprise and may even seem paradoxical at first. However, there might be a plausible explanation. News coverage of entrepreneurial events may in fact be lower in regions in which individuals indicate lower exposure. Again, it may be the difference between inputs and outputs into the journalistic processes that play a role here. Our press release data only cover the input, and it is unclear to what extent this is representative for the actual output, i.e., what can be found in newspapers. In this case, MEDIA reflects regional differences in the journalistic filtering process. However, there might also be another mechanism at work. The survey question underlying MEDIA actually aims at news coverage in national media, which may or may not reflect the situation in regional news. Respondents from regions with relatively high coverage of regional entrepreneurship activities may compare this coverage at the national level, which may cast a negative verdict about its intensity and vice versa. However, this interpretation needs further research and remains speculative at this stage.

After we have discussed the spatial distribution of entrepreneurship-related press releases and their relation to other entrepreneurship variables, we focus on the in which these press releases are expressed and what sentiments they contain. Table 3 represents the results of the regressions with the regionally aggregated sentiment scores of the entrepreneurship-related press releases as dependent. Columns one to five show the results of the OLS regressions that are estimated with heteroskedasticity robust standard errors. While all residual diagnostics (normal distribution, VIF, autocorrelation, homoskedasticity) confirm the appropriateness of the estimations, we run additional robustness checks. Firstly, we estimate the significance using a bootstrap approach. The corresponding upper and lower boundaries of the 95% confidence interval are given in columns 6 and 7. Secondly, we calculated a binary logistic regression with the dependent variable being one if the sentiments of entrepreneurship-related press releases are higher than on average across the regions. All models yield relatively comparable results, that is, most variables remain insignificant. While regional population (POP), population density (POP_DEN), and GDP become significant in the full OLS model (model 5), the according coefficients in the bootstrapped and logistic models remain insignificant. This casts doubts on their robustness. Nevertheless, with some caution, it suggests that entrepreneurship-related press releases are more positive in urban regions (high population density), while more negative in large regions (large population) and those that are economically better of (higher GDP per capita).

A robust result is obtained for the variable MEDIA. It is significantly positive in all models. At first this finding seems to contradict the regression results on the number of entrepreneurship-based press-releases (Table 2). However, here, it clearly corresponds to our expectations. Press releases in regions with higher MEDIA scores, i.e., higher levels of perceived media coverage of successful new businesses, contain more positive sentiments of entrepreneurship. Given the cross-sectional nature of our analysis, we cannot make a causal inference here. Nevertheless, the result suggests that frequent reporting about entrepreneurship and narratives about new businesses in the news, might be able to impact sentiment towards entrepreneurship at the regional level. In any case, the finding confirms a link between news based data and that obtained by surveys. Given that sentiments are shown to drive economic development in general (Baker et al. 2016), it can be expected that this relation also holds for regional entrepreneurial activities in particular. Accordingly, while our results only indicate that news may influence sentiments towards entrepreneurship, it can be argued that they are likely influencing actual entrepreneurship activities. While such interpretation is rather explorative and speculative at this stage, it clearly outlines new avenues for future research exploiting new(s) data and sentiment analyses.

This finding is in line with the (positive) display of role model entrepreneurs and new businesses. It also corresponds to the theory body on the impact of narratives on entrepreneurship and shows that journalist as senders within a sender-receiver model, can have significant influence on how phenomena are perceived.

In any case, to the best of our knowledge, the regression results are the first findings on systematic variations of topic-specific sentiments at the regional level. Moreover, they are also the first that explain parts of these variations, although admittedly, the parts are rather small. With these findings in mind we explored a second news data source, pressebox.de. Pressebox covers mainly news on technical and innovative content as soft and hardware, e-commerce and such. While showing similar patterns, the overall sentiment of this site (unrelated to news content) was much higher than the dpa source which prohibited merging (in addition to having a specific content emphasis). However, this second data set might yield some confirmation as well as further insight which we will explore in future research.

Table 2: Regression results

	negative binomial				*negative binomial*		*Poisson*	*Poisson*	*Poisson*
	General	General	General	General	Entrepreneur	Start up	General	Entrepreneur	Start up
RELEASES	0.0003***	0.0001***	0.0002***	0.0002***	0.0002	0.0003***	0.0001***	0.0002***	0.0002***
	(0.00003)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00001)	(0.00004)	(0.00001)
POP		0.0001***	0.0001***	0.0001***	0.0001**	0.0001*	0.0001***	0.0001***	0.00003***
		(0.00002)	(0.00002)	(0.00002)	(0.00004)	(0.00003)	(0.00000)	(0.00002)	(0.00001)
POPDEN			0.00005	-0.001**	-0.00003	-0.001	-0.0002***	-0.0002	-0.0005***
			(0.0003)	(0.0003)	(0.001)	(0.0004)	(0.00003)	(0.0002)	(0.0001)
GDP_PC				0.034***	0.017	0.021	0.040***	0.030***	0.068***
				(0.012)	(0.027)	(0.020)	(0.002)	(0.012)	(0.004)
STUDS				0.012***	0.011	0.005	0.012***	0.018***	0.004**
				(0.004)	(0.009)	(0.007)	(0.001)	(0.005)	(0.002)
UNEMPL			-0.039	0.034	-0.120	0.103	0.057***	-0.103	0.200***
			(0.054)	(0.056)	(0.136)	(0.093)	(0.011)	(0.077)	(0.026)
EAST	-0.039	0.036	0.240	0.150	1.644**	0.107	0.282***	1.441***	0.007
	(0.221)	(0.204)	(0.354)	(0.328)	(0.784)	(0.546)	(0.060)	(0.414)	(0.150)
FOUND	-0.053	-0.055	-0.058	-0.048	-0.117	-0.128**	-0.031***	-0.070	-0.105***
	(0.038)	(0.035)	(0.035)	(0.033)	(0.086)	(0.056)	(0.007)	(0.054)	(0.020)
MEDIA	-0.044**	-0.031*	-0.037**	-0.035**	-0.137***	-0.097***	-0.027***	-0.092***	-0.085***
	(0.018)	(0.017)	(0.018)	(0.017)	(0.047)	(0.030)	(0.004)	(0.027)	(0.009)
Constant	6.026***	4.877***	5.293***	3.635***	5.478**	5.099***	3.046***	2.816*	2.841***
	(0.873)	(0.850)	(0.993)	-1.050	-2.689	-1.760	(0.225)	-1.638	(0.542)
McFadden	0.11	0.13	0.13	0.14	0.18	0.17	0.88	0.74	0.86
Moran I (p-value)	0.18 (0)	0.16 (0)	0.14 (0.01)	0.16 (0)	-0.01 (0.5)	0.04 (0.23)	0.14 (0.01)	-0.02 (0.55)	0.06 (0.13)
Overdispersion p-value							3.1 (0)	1.99 (0.02)	2.32 (0.01)
Observations	96	96	96	96	96	96	96	96	96
Log Likelihood	-492.126	-482.930	-482.669	-474.032	-153.836	-310.674	-1,463.484	-182.589	-680.724
theta	1.883***	2.251***	2.262***	2.695***	0.886***	1.106***			
Akaike Inf. Crit.	994.253	977.859	981.337	968.064	327.672	641.347	2,946.968	385.179	1,381.448
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 3: Sentiment analysis

	Model 1	Model 2	Model 3	Model 4	Model 5	Lower CI 95%	Upper CI 95%	Logit, 1>mean(sentiment)
log(NO_ENTRE_SENTI+1)	-0.283	-0.125	-0.140	-0.132	-0.362	-0.362	-0.362	-5.904
	(0.525)	(0.636)	(0.636)	(0.636)	(0.679)			-3.654
log(ENTRE_COUNT)	0.062	0.092	0.095	0.104	0.074	0.074	0.074	0.764
	(0.067)	(0.075)	(0.075)	(0.082)	(0.082)			(0.525)
log(RELEASES)	0.076	0.081	0.075	0.068	0.044	0.044	0.044	0.310
	(0.098)	(0.099)	(0.099)	(0.099)	(0.099)			(0.470)
log(POP)	-0.161	-0.293*	-0.272	-0.277				-1.314
	(0.137)	(0.162)	(0.166)	(0.170)				(0.869)
log(POPDEN)		0.189**	0.157	0.151	0.061	0.061	0.061	0.920
		(0.087)	(0.102)	(0.101)	(0.084)			(0.663)
log(GDP_PC)		-0.729	-0.711	-0.720	-0.643	-0.643	-0.643	-2.545
		(0.452)	(0.447)	(0.458)	(0.436)			-2.234
log(STUDS)		0.020	0.021	0.024	0.006	0.006	0.006	-0.061
		(0.039)	(0.040)	(0.040)	(0.045)			(0.286)
log(UNEMPL)		-0.194	-0.122	-0.119	0.057	0.057	0.057	-0.181
		(0.129)	(0.170)	(0.173)	(0.175)			-1.136
EAST			-0.103	-0.089	-0.347	-0.347	-0.347	-2.219*
			(0.170)	(0.169)	(0.212)			-1.239
log(FOUND)				0.049	0.033	0.033	0.033	-0.043
				(0.121)	(0.124)			(0.697)
log(MEDIA)					0.955*	0.955	0.955	9.347**
					(0.519)			-3.634
Constant	1.150	3.918**	3.785*	3.805*	-1.936	-1.936	-1.936	-24.294
	(0.819)	-1.924	-1.915	-1.953	-2.115			-15.847
adj. R2	0.02	0.04	0.03	0.02	0.02			0.13
Breusch-Pagan	0.019	0.047	0.079	0.12	0.232			
Max VIF	4.9	5.4	5.4	5.7	5.4			5.4
Shapiro	0	0.006	0.009	0.011	0.019			
Moran I (p-value)	0.03 (0.28)	0 (0.43)	-0.01 (0.49)	-0.01 (0.5)	-0.05 (0.71)			-0.05 (0.73)

Note: *p<0.1; **p<0.05; p***p<0.01

5. Conclusion

We set out to explore a new dataset—media coverage in the form of news articles—to establish whether reporting of news on regional entrepreneurial events could be a potential new data source for entrepreneurship research and whether it can be characterised as Big Data. The discipline has become quite saturated over the recent decade in many areas by relying on surveys and registers, but new approaches and data sources are necessary to progress and to ultimately address fundamental issues in entrepreneurship research. We could show that different sets of (entrepreneurship-related) keywords lead to diverging results, indicating that content specific analysis of media coverage will be possible. Although the relationship between news coverage and early-stage entrepreneurial activity is rather weak, we are confident that data on news coverage could be used to unveil and differentiate different kinds of entrepreneurial activities and should be revisited. However, at this stage, our empirical results reject the idea of a strong relationship between actual entrepreneurial activities in regions and the intensity of it being reported.

By exploring the sentiment of entrepreneurship news, it became clear that different news sources discriminate systematically regarding the tone of their articles. This introduces challenges regarding merging and building a huge dataset of many news sources, but it also opens up many interesting questions. Sentiment analysis has furthermore shown quite interesting potentials regarding the evaluation of a qualitative dimension as an addition to a quantitative one.

The paper's contribution to the ongoing debate about the value of big data and internet-based information for entrepreneurship research is twofold. First, big data used in this paper has shown to be a serious option when looking for new indicators of regional entrepreneurial activity as news media, at least to a degree, covers such activities. Exploiting the potential of such data requires significantly less time, effort and capital than classical method to collect entrepreneurship data. Second, our approach is currently not able to serve as a complete substitute for traditional methods of data collection as they suffer from some important weaknesses in terms of data quality, accessibility and amount. A relevant contribution of our paper to the named debate is to show some of these limitations, despite the given potential. The combined analysis of big data sources with different analytical approaches seems to offer a completely new level of insight.

The results of this paper have to be interpreted in light of certain limitations. The reported news in the database are localised through algorithms that, although quite good, could be improved. For instance, it might be that a piece of news is related to multiple events in different locations, with entrepreneurship being just one of these events taking place in only one of these locations. Moreover, we have not yet explored the possibility of 'negating' news. Furthermore, our indicator of entrepreneurial activities is based on a survey and our data on news covering a very specific sub-section of all news, i.e., news that are perceived to have the potential of being reported but not necessarily those that are actually reported. So far, it is unclear to what extent our news data actually cover what is reported in national or local newspapers, social media or alternative outlets.

We also found a mismatch between actual reporting of entrepreneurial news and the perception of it, which opens up potential for further research on perception data versus reported events. However, sentiment analysis shows that in regions with higher perception of news about new successful businesses, the reporting on entrepreneurship-related news is more prone to be positive. This paper contributes towards developing new approaches to entrepreneurial phenomena. Although we cannot show a clear impact of regional entrepreneurial activity on regional entrepreneurial news reporting, these 'negative' results progress the knowledge base by contributing iterations that show the necessity of taking different paths or refining, updating and developing the used dataset. Those iterations are an integral part of approximation towards new knowledge.

Big Data approaches come with challenges of their own, as also pointed out by Fan et al. (2014), that are not necessarily comparable to problems traditional approaches had and have. This raises the need and opportunity for interdisciplinary research as the research becomes ever more computerised and complex. Despite some limitations common to work with new data sources, our empirical approach opens up a vast array of future research possibilities. This includes extending the data source to information on what event has found its way into the printed and electronic media; what has been covered in social media; and what receives 'just' regional and not national attention. This implies harvesting actual published news about entrepreneurship rather than just the potentially published material provided by the dpa.

In general, going forward, one goal has to be to broaden the current data basis of research and modernise it by adapting to data from recent, fast-paced sources such as dpa or pressebox. Further research should also explore the usefulness of other new possible Big Data sources for entrepreneurship research besides media coverage. However, news-coverage-based approaches such as ours should try to enhance the data basis by finding ways to reasonably merge existing news portals or feeds for increased regional coverage and broaden the empirical basis to open up even more options for analysis. New and untapped databases of articles, such as regional newspaper archives, need to be integrated to increase the amount and spread of regionalised data. Pursuing the question of whether the actual content of news articles has an impact or if it is simply based on the headlines of those articles since most overstimulated recipients of digital media only scan the news supply and just read what piques their interest through click-bait titles would be difficult, but scientifically very rewarding.

One other major challenge will be the determination and unravelling of causal linkages between news as a form of narratives and regional entrepreneurial activity or even entrepreneurial ecosystem conditions or development. As theory hints, it is clearly interdependent, to which extent and through which mechanisms have to be pursued empirically. How does which kind of entrepreneurial process coverage or narrative impact the regional start-up rate or quality? As new fields of data and data processing open up, the possibilities and need for new approaches and research questions emerge.

Entrepreneurship researchers may learn several lessons from this paper. First, while the used media data sources undoubtedly have large potential to at least partially replace current methods to collect data on entrepreneurial activity at the regional level, more experiences and results are needed to better assess the precise value of these new data sources. Also, in future research, scholars may regress alternative dependent variables than the GEM TEA rate as proxies for entrepreneurial activity. At least in Germany, other options do exist at the regional level. Second, while we have made a first step into a Big Data-based attempt to explore whether real entrepreneurial activities are covered by media news (i.e., by people who publish news about entrepreneurs and entrepreneurship), more knowledge about the precise mechanisms journalists, bloggers and other individuals employ to create this news is required. In that sense, it is important to explore the determinants affecting these individuals when they are writing the news. Contexts probably play an important role for these mechanisms, too, both from a person-related perspective and from a regional perspective.

Big (media) data based upon news have the potential to better cover both kinds of contexts (but the latter one in particular) than classically collected data. Similar to other academic disciplines, entrepreneurship research in general and spatially motivated entrepreneurship research in particular will profit from the disruptive effect of new digital technologies to collect and analyse Big Data. The geographical dimension and relevance of entrepreneurship is meanwhile widely acknowledged because it influences entrepreneurial attitudes as well as entrepreneurial success and consequently, the economic impact, output and outcome of entrepreneurship activities on economies. The regional environment is an important part of the overall context shaping entrepreneurial processes. This recent contextual turn in entrepreneurship research is obvious (see, e.g., Welter et al. 2019). An important issue of empirical research focussing on regional aspects, often in combination with micro-data with respect to the individual entrepreneur, is the lack of sufficiently large samples.

Our empirical research was inspired by some significant research gaps in entrepreneurship theory. Namely, the idea of regional entrepreneurial ecosystems as the rising star among theories of regional entrepreneurship suffers from adequate data sources to cover the complex and systemic relationships within such a regional system. These kinds of concepts generated in a deductive way of borrowing from innovation system literature and evolutionary economic geography require empirical tests. Primary data from surveys for these kind of tests are difficult to get, at least if they should meet criteria like statistical representativeness and interregional comparability. Another contribution to entrepreneurship theory can be expected with respect to the effects of entrepreneurial activities (as part of the regional context) on the publication behaviour of those individuals who produce big (media) data news about entrepreneurship in the region they work and live in. Also, the sender-receiver model, recently introduced to entrepreneurship research on role model effects, can be used to measure the impact of entrepreneurs as senders of signals of journalists or bloggers (writing and publishing news about regional entrepreneurship) as receivers of such signals in a given regional context. Applying news data as we have done in a very first and explorative attempt may, at least potentially, reduce some of these research gaps.

To address the potential of Big Data methods for exploring the interdependent relationship between entrepreneurship activities and media coverage, we propose the following elements as an agenda for future research:

- apply entrepreneurship news as a Big Data source for other countries than Germany;
- use other kinds of Big Data on media coverage than the one used in this paper;
- test for other definitions of entrepreneurial activities (besides TEA);
- add a check with qualitative data to confirm some of the findings with quantitative data (interviews with entrepreneurship news producers about their perceptions of entrepreneurship in a given region); and
- control for regional attributes like media landscape and start-up scene (size, development phase, communication behaviour and others).

References

Alvedalen, J. & Boschma, R. (2017): A critical review of entrepreneurial ecosystems research: Towards a future research agenda. *European Planning Studies*, 25, 887–903

Amodu, L., Ekanem, T. Yartey, D. Afolabi, & O. Oresanya, T. (2016). Media Coverage of Entrepreneurial Innovation as a Determinant of Sustainable Development in Nigeria. Conference Paper. 3rd International Conference on African Development Issues 2016.

Arenius, P. & Minniti, M. (2005). Perceptual variables and nascent entrepreneurship. *Small Business Economics*, 24(3): 233-247.

Audretsch, D. (2012). Entrepreneurship research. *Management Decision*, 50(5), 755-764, https://doi.org/10.1108/00251741211227384.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics 131(4), 1593–1636.

Balahur, A., & Steinberger, R. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceeding of WOMSA, 9.

Blood, D. J., & Phillips, P. C. (1995). Recession headline news, consumer sentiment, the state of the economy and presidential popularity: A time series analysis 1989-1993. *International Journal of Public Opinion Research* 7(1), 2-22.

Bosma, N., Coduras, A., Litovsky, Y., & Seaman, J. (2012). GEM Manual. Version 2012-9: May. GEM.

Bosma, N., & Sternberg, R. (2014). Entrepreneurship as an urban event? Empirical evidence from European cities. *Regional Studies*, 48(6), 1016-1033, DOI:10.1080/00343404.2014.904041.

Cameron, A.C., & Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics*, 46, 347–364.

Carroll, C. E., & McCombs, M. (2003). Agenda-setting effects of business news on the public's images and opinions about major corporations. *Corporate reputation review* 6(1), 36-46.

Chen, H. M., Schütz, R., Kazman, R., & Matthes, F. (2017). How Lufthansa Capitalized on Big Data for Business Model Renovation. *MIS Quarterly Executive*, 16(1).

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From Big Data to big impact. *MIS Quarterly: Management Information Systems*, 36(4), 1165-1188.

Coviello, N., & Jones, M. (2004). Methodological issues in international entrepreneurship research. *Journal of Business Venturing* 19 (2004), 485-508.

Coyne, C.J., & Leeson, P.T. (2004). Read All About it! Understanding the Role of Media in Economic Development. *KYKLOS*, 57(1), 21-44.

De Boef, S., & Kellstedt, P.M. (2004). The political (and economic) origins of consumer condence. *American Journal of Political Science* 48(4), 633-649.

Denzau, A. D. & North, D. C. (1994). Shared mental models: ideologies and institutions. *Kyklos*, 47: 3–31.

Dodge, M. & Kitchin, R. (2005). Codes of life: Identification codes and the machine-readable world. *Environment and Planning D: Society and Space* 23(6): 851–881.

Doms, M. E., & Morin, N.J. (2004). Consumer sentiment, the economy, and the news media. FRB of San Francisco Working Paper No. 2004-09.

Ebert, T., Götz, F.M., Obschonka, M., Zmigrod, L. & Rentfrow, P.J. (2019). Regional variation in courage and entrepreneurship: The contrasting role of courage for the emergence and survival of start-ups in the United States. *Journal of Personality*. 1–17.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, 1(2), 293-314.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Feldman, M.P. (2001), 'The Entrepreneurial Event Revisited: Firm Formation in a Regional Context'. *Industrial and Corporate Change* 10, 861–891.

Fogarty, B. J. (2005). Determining economic news coverage. International Journal of Public *Opinion Research* 17(2), 149-172.

Fritsch, M., Obschonka, M., Wyrwich, M., Gosling, S., Rentfrow, P. & Potter, J. (2018). Regionale Unterschiede der Verteilung von Personen mit unternehmerischem Persönlichkeitsprofil in Deutschland – ein Überblick (Regional differences of people with an entrepreneurial personality structure in Germany – An overview). *Raumforschung und Raumordnung* (Spatial Research and Planning), 76(1), 65-81.

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2016). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, https://doi.org/10.1111/ecin.12364

Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." Icwsm7.21 (2007): 219-222.

Goidel, R. K., & Langley, R.E. (1995). Media coverage of the economy and aggregate economic evaluations: Uncovering evidence of indirect media effects. *Political Research Quarterly* 48(2), 313-328.

Greenwood, B. N., & Gopal, A. (2017). "Ending the Mending Wall: Herding, Media Coverage, and Colocation in IT Entrepreneurship," *MIS Quarterly*, 41(3), 989-1007.

Hang, M., & Van Weezle, A. (2007). Media and entrepreneurship: A survey of the literature relating both concepts. *Journal of Media Business Studies*, 4(1), 51–70.

Hindle, K., & Klyver, K. (2007). Exploring the relationship between media coverage and participation in entrepreneurship: initial global evidence and research implications. International Entrepreneurship and Management *International Entrepreneurship and Management Journal*, 3(2), 217-242.

Isenberg, D. (2010). How to start an entrepreneurial revolution. *Harvard Business Review*, 88(6), 40–50.

Kitchin, R. & McArdle, G.A. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society* 3 (1), DOI: 10.1177/2053951716631130.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining Big Data to extract patterns and predict real-life outcomes. *Psychological Methods*, 21(4), 493.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. In: Meta Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Mahmoodi, J., Leckelt, M., van Zalk, M. W., Geukes, K., & Back, M. D. (2017). Big Data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, 18, 57-62.

Malecki, E.J. (2018). Entrepreneurship and entrepreneurial ecosystems. Geography Compass, 12, DOI: 10.1111/gec3.12359.

Marr, B. (2014). Big data: The 5 vs everyone must know. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.

Marz, N. & Warren, J. (2012). Big Data: Principles and Best Practices of Scalable Realtime Data Systems. MEAP edition. Westhampton, NJ: Manning.

Mayer-Schonberger, V. & Cukier, K. (2013). Big Data: A Revolution that will Change How We Live, Work and Think. London: John Murray.

Obschonka, M. (2017). The quest for the entrepreneurial culture: psychological Big Data in entrepreneurship research. *Current Opinion in Behavioral Sciences*, 18, 69-74.

Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet.

Presseportal (2018). https://www.presseportal.de/about. Accessed 31 May 2018.

Remus, R., Quasthoff, U., & Heyer, G. (2010, May). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In LREC.

Richard, A., & Gall, J. (2017). A bag-of-words equivalent recurrent neural network for action recognition. Computer Vision and Image Understanding, 156, 79-91.

Roundy, P.T. (2016). Start-up Community Narratives: The Discursive Construction of Entrepreneurial Ecosystems. *Journal of Entrepreneurship* 25 (2): 232-248.

Shiller, R. (2017). Narrative economics. *American Economic Review* 107: 967-1004.

Spigel, B. (2015). The relational organization of entrepreneurial ecosystems. Entrepreneurship Theory and Practice. DOI: 10.1111/etap.12167.

Feuerriegel, S. & Proellochs, N. (2018). SentimentAnalysis: Dictionary-Based Sentiment Analysis. R package version 1.3-2. https://CRAN.R-project.org/package=SentimentAnalysis:

Sternberg, R. (2009). Regional Dimensions of Entrepreneurship. Boston, Delft: Now Publishers (= *Foundations and Trends in Entrepreneurship*, 5(4)).

Sternberg, R., Wallisch, M., Gorynia-Pfeffer, N., von Bloh, J., & Baharian, A. (2018). Global Entrepreneurship Monitor (GEM). *Länderbericht Deutschland* 2017/18 (=National Report Germany 2017/2018). Eschborn and Hannover; RKW Kompetenzzentrum and Institute of Economic and Cultural Geography, Leibniz University Hannover.

Stuetzer, M., Obschonka, M., Brixy, U., Sternberg, R. & Cantner, U. (2014). Regional characteristics, opportunity perception and entrepreneurial activities. *Small Business Economics* 42(2), 221-244.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

Wang, F., Mack, E. A., & Maciewjewski, R. (2017). Analyzing Entrepreneurial Social Networks with Big Data. *Annals of the American Association of Geographers*, 107(1), 130-150.

Wartick, S. L. (1992). The relationship between intense media exposure and change in corporate reputation. *Business & Society* 31(1), 33-49.

Welter, F., Baker, T. & Wirsching, K. (2019). Three waves and counting: the rising tide of contextualization in entrepreneurship research. *Small Business Economics* 52 (2), 319-330.

Welter, F., Baker, T., Audretsch, D., & Gartner, W. (2017). Everyday Entrepreneurship—A Call for Entrepreneurship Research to Embrace Entrepreneurial Diversity. *Entrepreneurship, theory and practice* 41(3), 11-321.

Wyrwich, M., Stuetzer, M. & Sternberg, R. (2016). Entrepreneurial role models, fear of failure, and institutional approval of entrepreneurship: a tale of two regions. *Small Business Economics* 46(3), 467-492.

Wyrwich, M., Sternberg, R. & Stuetzer, M. (2018). Failing role models and the formation of fear of entrepreneurial failure: a study of regional peer effects in German regions. *Journal of Economic Geography*. Online First: https://doi.org/10.1093/jeg/lby023