Linköping University | Department of Management and Engineering

Master's thesis, 30 ECTS | Computational Social Science

2021 | LIU-IEI-FIL-A--21/03705--SE

Blank spots, critical information needs and local journalism funding

Measuring the impact of state financial support for local journalism on the critical information needs of citizens in Swedish journalistically fragile municipalities

Simona Bisiani

Supervisor : Benjamin Jarvis, Étienne Ollion
Examiner : Karl Wennberg

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Copyright

Abstract

A global business model crisis in journalism, fuelled by loss in advertising revenue, challenges the survival of local news production. In Sweden, it has led to the closure of several newspapers across the country, and the concentration of ownership of local news outlets in the hand of a few, remotely located, news media organizations. This phenomenon has given life to what researchers in Sweden have named *blank spots*, municipalities lacking an editorial presence within them. Meanwhile, research has emerged indicating both the importance of local newspapers in providing critical information to their communities, and empirical differences in journalism between blank spots and municipalities with an editorial presence. In 2019, to facilitate, encourage and sustain local journalism, the Swedish government introduced financial support for outlets producing local news in fragile municipalities, no matter the ownership or the production location. This research investigates the impact of funding on local news content in several Swedish blank spots. Particularly, it does so by aggregating articles in categories that describe the type of information they provide to communities, in line with what has emerged in research as the Critical Information Need framework. Pioneered by US researchers, it measures the state of local journalism by quantifying the relevance of a news piece for a citizen to operate functionally within a democratic system. This research is the first in Sweden to adopt the CINs framework, and it avails of computational text analysis to do so, introducing a novel methodological approach to the literature on local journalism and blank spots in Sweden. Using Facebook's FastText algorithm for supervised text classification, the CIN category of 51209 articles about six blank spot municipalities was predicted. The results show that the majority of articles published in the sampled municipalities fulfills a critical information need, suggesting that blank spots, while not having locally present editorial offices, are still being provided with relevant news. Finally, a statistical analysis is carried out to measure the impact of funding on the quantity of CINs news articles, and articles about civic information and politics, for a funding receiving publisher, as well as the municipality he is embedded in. The results show a mixed picture of no significant changes for the funded publishers, yet somewhat of a positive picture for the municipalities those publishers are embedded in. The results of the analysis complement the literature understanding of the funding as having had a mostly invisible, and otherwise marginally positive, effect on the volume of articles in Swedish journalistically fragile municipalities.

Acknowledgments

I would like to thank my supervisors, Benjamin Jarvis and Étienne Ollion, for the invaluable input and encouragement during the creation and writing of this thesis. Particularly, I would like to thank Benjamin Jarvis for his support in the sampling strategy and the statistical side of this thesis, and Étienne Ollion for his dedication in helping me with the classification task.

Furthermore, I would like to thank my classmate Pablo Bello Delpón for lending me an ear and his precious time, and discussing, with interest, the workings of my project; for telling me about Concept Mover's Distance and showing me how to use it, thus allowing me to learn something new. Similarly, I would like to thank my coursemate Abigail Galea for encouraging me when I needed it the most, and for her willingness to listen, discuss, and advise on many different aspects of this thesis.

Last, but not least, I would like to thank my family and friends, that in various ways have helped me throughout the time of the project.

Contents

List of Figures

List of Tables

1 Introduction

Journalism, worldwide, is subject to rapid changes, due to evolving technology, increasing competition, and adapting business models. Digital technologies have substantially altered the form of journalism and opened the path to new alternatives to the press, television, and radio. Simultaneously, the Internet has given new attractive avenues to advertisers, causing financial distress to commercial news providers, which for decades used advertising revenue as their major source of income (Abernathy, 2018). Attempts to shift to alternative income streams such as subscription models and digital advertising have so far failed to make up for the lost revenue (Miller, 2018). In the US, this business model crisis has caused, since 2004, the closure of one in five newspapers and reduced the number of journalists by half (Abernathy, 2018). While new forms of journalism are being established, such as online-only news outlets and so-called citizen journalism, the closure of newspapers is still a cause for concern (Abernathy, 2018; Mahone, Wang, Napoli, Weber, & McCollough, 2019; Miller, 2018). On the one hand, local press reporters are the producers of most stories about local institutions, politics and corruption, acting as watchdogs of the local communities. Furthermore, there is a growing understanding that newspapers, particularly local ones, are the medium fulfilling communities' Critical Information Needs, understood as a set of eight categories of information that enable citizens to function in a democratic system, such as information about politics, civic life, environmental threats, economic opportunities, and more (Abernathy, 2018; Mahone, Wang, Napoli, Weber, & McCollough, 2019; Waldman, 2011). As local newspapers are vanishing, scholars have started monitoring the number and geographical distribution of the surviving papers. In the US, communities lacking a local newspaper have been given the name of *news deserts* (Abernathy, 2018).

Similarly, Swedish researchers have identified several *vita fläckar*, or *blank spots*, municipalities lacking an editorial office due to closure of newspapers or centralization of news production (Nygren & Schjærff Engelbrecht, 2018). In Sweden, concerns around weakened local journalism have culminated in the creation of a new form of state financial support delivered to news outlets to help them sustain the production of local news in blank spots or otherwise journalistically fragile municipalities. Thirty million SEK were distributed in 2019 to 34 publishers across 65 municipalities, which increased to 121 million SEK and 90 publishers across 181 municipalities in 2020 (Nygren, 2020b). More municipal officials within those municipalities where resources were distributed see a positive, rather than negative, effect on local journalism, although around half report no great difference (Tenor, 2020).

Beyond the survey, however, there is no knowledge yet on how the funding has impacted local journalism. Furthermore, existing research on blank spots has so far focused on structural aspects of local journalism, such as ownership of newspapers and differences in the number of outlets across municipalities of various type and size (Nygren, 2020b; Nygren & Nord, 2020). The Swedish literature lacks, to this day, a systematic assessment of local journalism in blank spots in relation to citizen's vital information needs. Thus a question remains unanswered: are local newspapers in blank spots, often produced from afar, providing citizens

with that vital local coverage that newspapers have been deemed to generate? In this research I identify in Mahone, Wang, Napoli, Weber, and McCollough (2019)'s notion of Critical Information Needs (CINs) a framework that would help provide a different understanding of local journalism in Sweden, one that focuses on local newspapers' role as producers of vital local information. This research thus imports the notion of critical information needs to Sweden, and particularly quantifies how local newspapers in blank spots serve citizens' critical information needs. To do so, it will avail of methodological tools that draw from the field of computational social science, introducing novel methods to the literature on blank spots and local journalism in Sweden. Finally, this study delivers an assessment of the impact of the newly introduced state funding on the fulfillment of blank spots citizens' CINs through their local newspapers, looking particularly at the type of locally relevant news that outlets have traditionally been accounted to provide: civic and political articles.

1.1 Research questions

The research questions that are expected to be answered are:

- *RQ1: are local newspapers in blank spot municipalities providing articles that fulfill the critical information needs of their citizenry?*
- *RQ2: how has the newly introduced state funding impacted local journalism' provision of critical information needs at the publisher and municipality level, and particularly in terms of the type of stories local newspapers have traditionally been associated with, namely civic and political stories?*

1.2 Overview of the research structure

This study uses three continuous years of news articles covering six comparable blank spot municipalities in Sweden, three which have received funding and three which have not. A supervised machine learning model was trained to code whether the articles served a critical information need, and a statistical analysis was carried out to assess differences in changes in the critical information needs before and after treatment, at the publisher and municipality level, between treated and untreated subjects.

To implement this project, different data sources were brought together. The identification of blank spots was done using *Kommundatabasen*, a website supported by the Intitutet för Mediestudier, whose purpose is to keep track of the newsrooms in Sweden, their ownership status and the municipality their office is present in (Institutet för Mediestudier, 2021). Information about funding was retrieved by publicly available reports on the Myndigheten för Press, Radio och TV (MPRT) website, specifying which news outlets have received funding, for which year, and how much money was allocated. Having identified only a few blank spots not having received any funding, I matched a sample of those with their nearest blank spot neighbour, based on demographic correlates of news production and consumption. This matching process was carried out on the website Kolada, an open database from RKA, a non-profit organization that promotes tools to do comparative studies of Swedish municipalities. News articles for the sampled pairs of blank spot municipalities were collected through Retriever, also known as MedieArkivet, one of the most comprehensive web archives for news articles produced in the Nordic region.

A chart of the conceptual framework is presented in Figure 1.1.

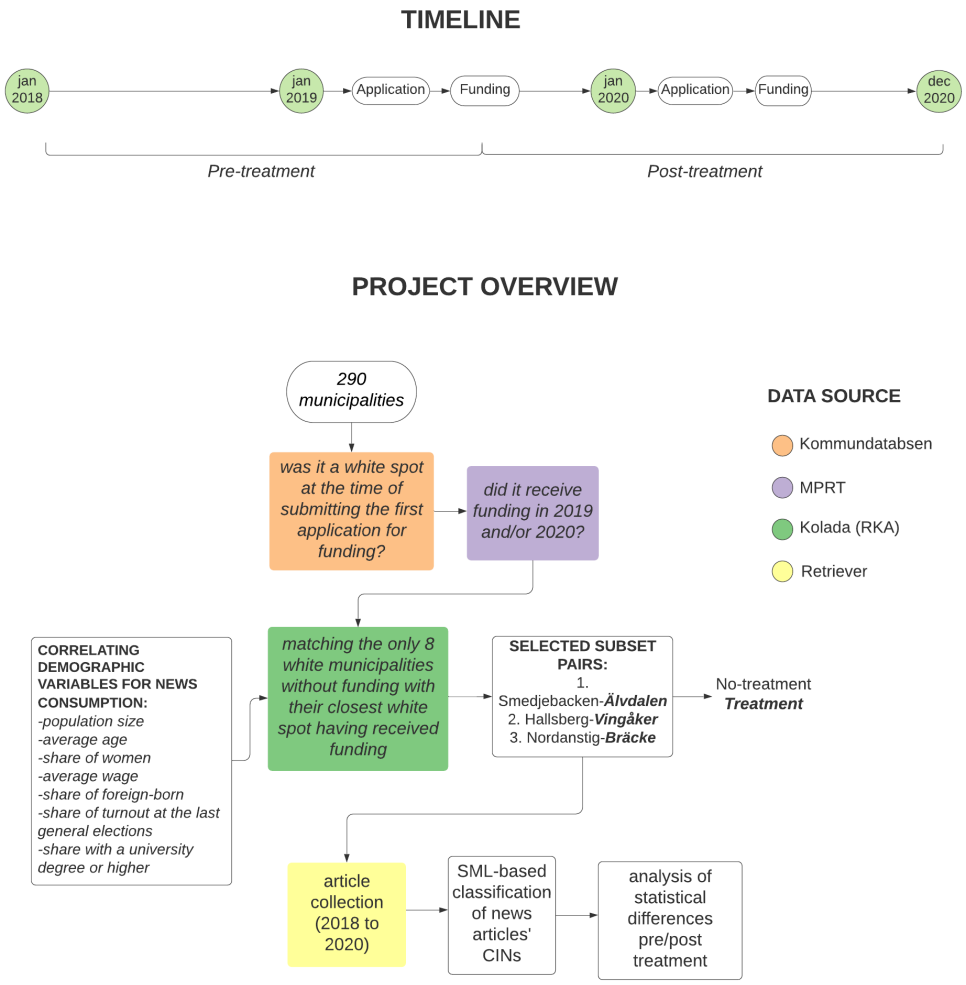


Figure 1.1: A chart of the research structure.

The text classification, done through a supervised machine learning model, shed light on the type of information provided by local newspapers. The results show that around 75% of articles published in the sampled blank spot municipalities fulfill a critical information need, regardless of funding allocation. The patterns are consistent across the different blank spots, while reflecting all the same the inherent characteristics of the municipalities or the occurrence of extraordinary events that demand local coverage attention, such as the Swedish wildfires of 2018, the General Elections of 2018, and the Covid-19 pandemic in 2020. Overall, the classification confirms that local newspapers produce important information for their communities. The statistical analysis revealed that the funding did not lead to a significant increase in the number of articles about civic information and politics, which historically have been associated with the figure of the local press reporter. However, the funding seemed to have had a small, yet positive and significant, effect on the volume of articles in the municipalities where some of its publishers have received funding, and that this holds when excluding less relevant stories such as obituaries, sport, and other. However, the effect is only found when measuring the impact on the journalistic production of the municipality as a whole,

and not when only looking at the production of the funded publishers, suggesting perhaps the idea of a crowding out effect of the funding. Overall, the results offer some guidance on the fact that the recently introduced funding has, whenever noticeable, only marginally positively altered local journalistic content, by means of increasing the volume of articles about a CIN in the municipalities in which some publishers have obtained the subsidy. A question thus emerges in regards to the alternative ways in which it might have alleviated outlets' pressures, and whether any stronger effect might occur in the long term.

1.3 Outline

The structure of the thesis is as follows: in Chapter 2, an introduction to key concepts and relevant literature is presented. In Chapter 3, an overview of the data and method is given. In Chapter 4, the analysis' results are shown. Chapter 5 discusses the results and points to the limitations of this study, while fitting this thesis' findings into the wider literature. Lastly, Chapter 6 summarises the scope and findings of this research and points to possible extensions for future research.

2 Literature Review

2.1 The digital revolution's impact on newspapers

The digital revolution has radically transformed society, introducing dynamic modes of sourcing, displaying and sharing information. For the field of journalism, this has implied profound changes, the implications of which are currently unfolding. In particular, there have been drastic changes in the news ecosystem (Wihbey, 2019). There is now a greater and more varied number of actors partaking in the news-making process, but simultaneously the survival of traditional news agents is being threatened by a business model crisis, stemming from technological advancements. Particularly, we are witnessing internationally a reduction in the number of local newspapers and local press journalists, the result of the struggle to sustain costs, as income shrinks due to advertisers shifting towards online websites instead (Miller, 2018).

In the US, 1800 local newspapers have disappeared since 2004, out of around 9000 publications. Additionally, the number of journalists has been reduced by half, leaving behind smaller newsrooms with less human resources to generate coverage, and less ambitious mission statements (Abernathy, 2018). In Sweden, the downsizing trend started in the late 1980s. Since then, nearly half of advertising revenue has disappeared, alongside around 40 percent of circulation, despite the country having a strong tradition of local press readership (Nygren, 2020a; Ohlsson, 2017). Just in the last fifteen years, the number of journalists in newspapers has reduced by around 25 percent, and approximately a third of all local newsrooms have closed down (Nygren & Althén, 2014).

2.2 News deserts and blank spots

Alarming closing rates of newspapers and shrinkage of newsrooms have led researchers internationally to start mapping journalistic presence across their countries. A comprehensive report on the state of journalism in the US points to 200 communities lacking a newspaper, no matter the release frequency, and an additional 1449 counties with just one newspaper, often a weekly (Abernathy, 2018). Following Stites (2011)' description in a NiemanLab article about the troublesome effects of cutbacks in journalism, these journalistically weak areas have been named *news deserts*.

In Sweden, researchers at the Intitutet för Mediestudier have been mapping the presence of newsrooms across the country since 2015 (Wallentin, 2020). In Sweden there are 290 municipalities, and all of them receive some sort of coverage. However, not all of them have an editorial office physically present within them, which indicates that the local news is in those cases produced from afar (Nygren & Schjærff Engelbrecht, 2018). The Swedish equivalent of a

news desert is thus not a municipality lacking a local newspaper, but rather the local presence of journalists and editors. Those municipalities have been named *vita fläckar*, or *blank spots*, and tend to occur in rural municipalities or in the outskirts of big cities (Nygren & Schjærff Engelbrecht, 2018). As Nygren and Althén (2014) explain, financial challenges have led to an increased concentration of media ownership in Sweden. Many small local newspapers have been bought by large media organizations, the eight biggest of which now own 80 percent of all daily newspapers. To minimise costs, many of these local news offices have been relocated to centralized production facilities, generating a debate over the potential detrimental effects of not having a locally based news office (Nygren, 2020a). One effect of the physical presence of editors and journalists within a local municipality has been shown to be a higher production of meaningful local stories (Pålsson, 2017). Similarly, remotely-operated journalism meant an increase in easy to produce news such as press releases, crime and accidents, the likely result of the fact that sources like the police and municipal websites are easy to access remotely (Hellekant Rowe, 2016). Beyond an editorial office, it is the presence of journalists living in a municipality, rather than having a locally present newsroom, that positively impacts readership rates (Kekezi and Mellander, 2017). Finally, a lack of an editorial office within the municipality indicates 10 percent less readership compared to municipalities with one within their borders (Nygren & Schjærff Engelbrecht, 2018). It appears then that the existence of blank spots impacts the relevance of journalism and readership rates in those municipalities, raising questions about whether their citizens are receiving important news about their local communities.

2.3 The Critical Information Needs framework

The vanishing of local newspapers raises the question of what is being taken down with them. Whether the closure of editorial offices and the loss of local reporters is at all an issue requires us to understand the role of local news and the news ecosystem at large. In the US, a major effort in this regard is represented by a 2011 investigation launched by the Federal Communications Commission (FCC) (Waldman, 2011). Triggered by Abernathy's (2018) research on news deserts, the FCC generated a series of reports, in 2009 and in 2012, aimed at understanding citizens' information needs, and how they are being met in the broadband age. The 2012 report, in particular, focused on community-level news media ecology, and highlighted the importance, for a community, to have its *critical information needs* (CINs) met.

The report states (Friedland, Napoli, Ognyanova, Weil & Wilson III, 2012:v):

> There is an identifiable set of basic information needs that individuals need met to navigate everyday life, and that communities need to have met in order to thrive. [...] Critical information needs [...] are those forms of information that are necessary for citizens and community members to [...] fully participate in the civic and democratic lives of their communities should they choose. To meet these needs, communities need access to the following eight categories of essential information, in a timely manner, in an interpretable language, and via media that are reasonably accessible, including information about:
>
> 1. *emergencies and risks*, both immediate and long term;
> 2. *health and welfare*, including specifically local health information as well as group specific health information where it exists;
> 3. *education*, including the quality of local schools and choices available to parents;

4. *transportation*, including available alternatives, costs, and schedules;
5. *economic opportunities*, including job information, job training, and small business assistance;
6. *the environment*, including air and water quality and access to recreation;
7. *civic information*, including the availability of civic institutions and opportunities to associate with others;
8. *political information*, including information about candidates at all relevant levels of local governance, and about relevant public policy initiatives affecting communities and neighborhoods.

Following the investigation, research emerged using the CINs notion in relation to local journalism. A study shows that across 100 randomly sampled US communities, around a tenth of those either did not have any articles serving a critical information need, or produced no original stories, while two in ten did not contain any local news stories (Napoli, Weber, MCcollough, & Wang, 2018). Furthermore, local newspapers have been shown to hold a vital informative role for communities, upon categorizing whether stories, across the overall media ecosystem, fulfilled a critical information need (Friedland, Napoli, Ognyanova, Weil, & Wilson III, 2012; Mahone, Wang, Napoli, Weber, & McCollough, 2019). There is now a general understanding that local newspapers act as pillars of local democracy by producing independent reporting that covers local affairs (Abernathy, 2018; Mahone, Wang, Napoli, Weber, & McCollough, 2019; Miller, 2018). Traditionally, most beat reporting, surveillance of local institutions, coverage of courts, local politics and cultural events, appears in local newspapers, and other news formats (e.g. television, radio) outsource this type of reporting from newspapers (Mahone, Wang, Napoli, Weber, & McCollough, 2019). Loss of newspapers thus translates to a loss in original news, which weakens the watchdog role journalists have traditionally held. Additionally, it is often from the type of journalism local newspapers make that the biggest investigative journalism stories have historically come out (Miller, 2018). Analogously, the closure of newspapers in the US has caused local governments to work less efficiently, perhaps precisely due to lack of scrutiny by the local newspaper (Gao, Lee, & Murphy, 2019). Within newspapers then, a void of meaningfully informative local news is forming, especially as those few remaining local journalists are asked to produce more stories that are sensational in nature (Downie Jr & Schudson, 2009). Consequently, the figure of the local reporter producing in-depth investigations unveiling tucked away information, misuse of financial resources by public institutions and exploring local social issues, is disappearing (Waldman, 2011). Finally, researchers have pointed out that news deserts tend to emerge in rural communities or in the outskirts of metropolitan areas, correlating with lower education levels, lower income, and an older population (Abernathy, 2018; Grant, 2020). It seems then that when it comes to being served equally by means of vital information, some groups are clearly disadvantaged, those that perhaps can afford it the least (Abernathy, 2018).

Can other media types such as local TV news and participatory journalism make up for the disappearance of newspapers? The gap does not seem to be filled by other players in the media ecological system. For example, the amount of reporting TV does on local news such as elections, education, healthcare and civic life is limited, while sensational news genres such as crime account for larger portions of the TV sendings (Kaplan & Hale, 2010). In Sweden, citizen journalism often does not emerge in the areas that are most weakly covered by newspapers, but rather on the side of legacy media, despite the aim for most being strengthening the local identity. Furthermore, often citizen journalism and legacy media contribute differently to the news ecology, and have different news standards and objectives (Nygren, Leckner, & Tenor, 2018). To conclude, both in the US and in Sweden research has shown that the closure of newspapers is leaving a gap in the news ecological system, undermining

citizens' opportunity to be informed on what happens in their communities, and losing that watchdog function that journalists have traditionally held.

However, the notion of critical information needs has yet to be incorporated in Swedish media studies. The CINs framework would provide a helpful basis to map the current state of local journalism, due to the understood importance of newspapers and the inability of hyperlocals to be equally credible or strategically located. The literature on blank spots has so far focused on structural aspects of local journalism and the news ecosystem, such as media ownership, the geographical distribution of blank spots, the role of citizen journalists, and the measurable differences between having a local editorial office within the municipality borders (Nygren & Althén, 2014; Nygren, Leckner, & Tenor, 2018; Nygren & Schjærff Engelbrecht, 2018). As suggested by Nygren and Schjærff Engelbrecht (2018) the CINs framework would offer a novel way to map the state of local journalism in Sweden, one that ties together local news content with the degree to which articles serve the community's CINs.

This research introduces the CINs framework to Swedish media literature. In particular, it looks at municipalities without an editorial office, and quantifies how the local journalistic coverage looks like from a CINs perspective, relating to the first research question *RQ1: are local newspapers in blank spot municipalities providing articles that fulfill the critical information needs of their citizenry?* Given this attempt is the first in the country, it is challenging to derive any meaningful hypotheses. Comparison to prior findings (see Weber, Andringa, and Napoli (2019)) should be made with caution, due to differences in the classification method, in journalistic form, in the definition of a news desert or a blank spot, as well as cultural context, between the Swedish and American communities. Rather, this question should be understood in light of the importance of measuring the state of local journalism in journalistically fragile municipalities from the perspective of the type of information it is being delivered to citizens through their local papers.

2.4 Lokaljournalistik stödet: the funding for local journalism

Concerns around the existence of blank spots, the closure of newsrooms, and the concentration of media ownership in Sweden has culminated in the creation of a new form of yearly state funding to support local journalism, known as *lokaljournalistik stödet*, and supplied by the Myndigheten för Press, Radio och TV (MPRT) (MPRT, 2019). The financial support was introduced in 2019, and it allocates a sum of money of the national governmental budget for media to news outlets dedicated to the production of local journalism. News outlets fulfilling a set of basic requirements (see Appendix A), were considered for the funding. The application had to be submitted by interested news outlets, in February 2019 and February 2020. The funding has been delivered in the summer of 2019 and of 2020, to those outlets whose application was approved by the MPRT, on the basis of the news outlet's characteristics, the outlet's proposition on how to use the funds, and the state of local journalism in the targeted municipality. Allocated funds have since been utilized in various ways. In some municipalities they have been used to simply sustain existing costs, or to hire a local reporter, whereas in Bräcke a new editorial newsroom has been opened within the municipality, which had lacked one since 2013 (Wallentin, 2020).

In the first year of the subsidy allocation, thirty million SEK were distributed to 65 municipalities. The figure increased to 121 million SEK and 181 municipalities in 2020 (Nygren, 2020b), indicating that many news outlets qualify for funding and that many municipalities qualify as journalistically fragile. It is yet to be assessed whether the funding has had any impact of the state of local journalism. A recent survey has shown that when asked if

they considered the coverage of local affair improved in those municipalities where resources were distributed, more municipal officials saw a positive, rather than negative, effect on local journalism, although around half of the respondents report no noticeable difference (Tenor, 2020). This research attempts to make a first assessment on whether we see any changes in the quantity and type of news blank spots receive and the funded publishers produce thanks to the subsidy. It is not a straightforward question whether the amount of journalism and its content composition have improved as a result of more financial resources directly aimed at helping the sustenance of local journalism. For example, because it is hard to hypothesize a direct improvement. Usage of funds varies across news outlets, and potential lags can be expected. Second, because increasing challenges in mere survival of a title could mean funding simply helps a newspaper stay alive, without generating any changes in the local journalism. Third, 2020 news coverage trends were impacted by the Covid-19 pandemic, which has affected media production and consumption (Nygren, 2020b). Unveiling the effect of funding on the relevance of news articles supplied to fragile municipalities would generate insights that can guide policy-makers on the outcome the financial measure to sustain journalistically fragile municipalities has had on the journalism itself. Simultaneously, we can gain an understanding of the degree by which critical information needs for Swedish journalistically fragile municipalities are fulfilled.

The second research question, pertinent to the funding, is *RQ2: how has the newly introduced state funding impacted local journalism' provision of critical information needs at the publisher and municipality level, and particularly in terms of civic and political stories?* Given that the aim of the funding is to sustain and strengthen local journalism, one plausible outcome is that more informative journalism is produced as a result of the financial support. This is formulated in the first hypothesis:
H1: the allocation of funding will have increased the volume of critical information needs articles in those municipalities within which some publishers have obtained funding, and particularly for those funding-receiving publishers.
Beyond an increase in volume, one relevant aspect to capture would be whether the funding has had an impact of the type of stories published. For instance, we could expect the funding to have incentivised the production of those type of articles that local press reporters have been traditionally deemed to provide, namely articles about city halls, corruption in politics, misuse of power, and keeping institutions accountable. These type of news could be generalised as belonging to the CINs categories of Civic Information and Politics. I formulate a second hypothesis:
H2: the allocation of funding will have increased the overall amount of articles about Civic Information and Politics.

2.5 Computational text analysis

Recent advances in computational power and technology have enabled the development of tools capable of conducting large-scale analysis of text. The task of analysing text has traditionally seen researchers manually examining documents to interpret them. Compared to hand-coding, the novel practice of automated text classification requires significantly less time, all while enabling the processing of large amounts of text. Analogously, efforts to preserve online and textual content has led to the creation of digital archives, creating a wealth of available textual content for researchers (Barbera, Boydstun, Linn, McMahon, & Nagler, 2021). While computational social scientists have been at the forefront of adopting tools from Natural Language Processing in their research, journalism researchers are only recently picking up on those (De Grove, Boghe, & De Marez, 2020; Lazer, Pentland, Adamic, Aral, Barabasi, Brewer, Christakis, Contractor, Fowler, Gutmann, et al., 2009). Some recent efforts

include Burggraaff and Trilling (2020)'s automated content analysis of just over 750000 news articles to identify differences in news values between online and offline articles, or Jacobi, Van Atteveldt, and Welbers (2016)'s usage of topic modelling to analyse trends in the New York Times' nuclear coverage. Recently, an attempt to use a supervised dictionary-based application to specifically code the eight critical information needs was conducted by Weber, Andringa, and Napoli (2019), who used the data from a Napoli, Weber, MCcollough, and Wang (2018) study where the critical information needs were hand-coded to build a dictionary of CIN-specific terminology. In Sweden, local journalism researchers are yet to use automated content analysis, and have so far relied on hand-coding of texts (Andersson, 2013; Hellekant Rowe, 2016; Nygren & Schjærff Engelbrecht, 2018; Truedson & Karlsson, 2018). In this research, I identify in supervised machine learning (SML) a viable tool to enlarge the scope of the project, while demonstrating the viability of using computational text analysis to measure the state of local journalism in Sweden. Particularly, I use Facebook's FastText model for text classification, leveraging on n-grams vector representations to preserve semantic meaning and structure for the prediction task (Joulin, Grave, Bojanowski, & Mikolov, 2016).

3 Data and Method

3.1 Pre-data collection steps

This thesis consists of a series of methodological steps. First, the identification of blank spots and the retrieval of information on funding was carried out. Second, a sample of blank spots to use for this study was selected. Third, news articles were collected, from the 1st January 2018 to the 31st December 2020, for six blank spot municipalities using the news media archive Retriever. Fourth, a SML model was built to classify whether an article fulfills a CIN. Lastly, I used difference-in-differences regression to test my two hypotheses about the impact of funding on the volume of articles about CINs and about civic information and politics.

3.1.1 The identification of blank spots

Following Nygren and Schjærff Engelbrecht (2018)'s definition, a blank spot is a municipality with no editorial newsroom physically present within it. The definition implies that one municipality can, across a longer period of time, gain the status of blank spot, lose it, or both. The turning of a municipality into a blank spot could be determined either by a local news outlet closing, or the relocating of an editorial office elsewhere, perhaps due to change in ownership. The identification of blank spots is made possible through to the Institutet för Mediestudier's *Kommundatabasen* website, aimed at keeping track of Swedish newsrooms. The website provides a list of newsrooms, and includes information such as the opening and closing dates of editorial offices, and the municipality the news outlet is based in. To note, it does not show which municipalities an outlet covers, but simply where it is located. In fact, there is not a system in place in Sweden to know which geographical area a particular newspaper covers, regardless of where their office is based. To appear in the Kommundatabasen website, newsrooms fulfill a set of criteria, such as publication frequency (see Appendix A for a detailed explanation of the criteria). For the purpose of this research, a municipality has been defined as a blank spot if it was one at the time of submitting an application for the first round of funding for local journalism, in early 2019. The reason behind this choice is that those municipalities would reasonably be the most journalistically fragile in the pre-treatment phase, and perhaps been prioritized for funding. By availing of a dataset from Kommundatabasen with all existent newsrooms in Sweden, their start date and their end date, whenever applicable, I filtered out for all newsrooms existent in early 2019, and, by observing where their offices were located, generated a mapping of the municipalities lacking an editorial office within them. A total of 45 blank spots municipalities were identified, which can be seen in Figure 3.1.

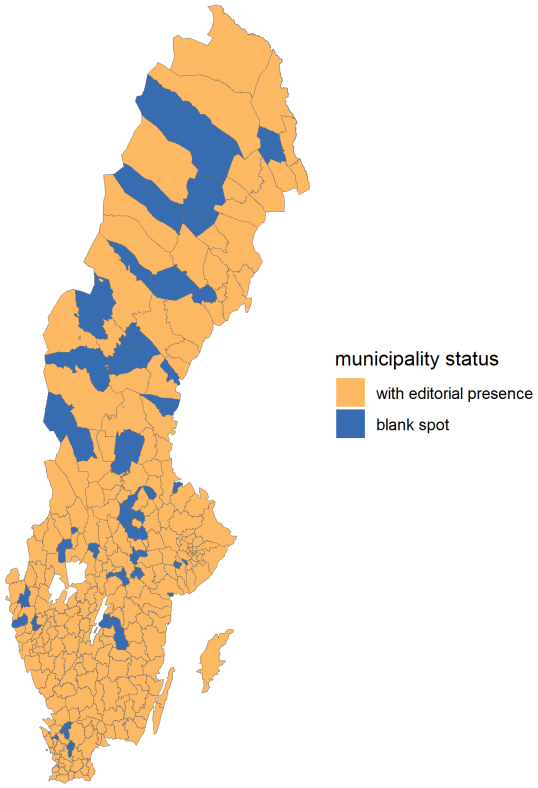


Figure 3.1: Geographical distribution of blank spot municipalities in January 2018. Data from Kommundatabasen

3.1.2 Who has received funding?

Information on funding allocation was publicly available on the MPRT website, and it comprised two lists, one for 2019 and one for 2020, of all news outlets which have received funding and the corresponding amount, as well the municipality for which they obtained funding for. To note, there is no information of funding applicants, but only of recipients, which suggests that possibly more newsrooms might have submitted applications for funding but were turned down. The turning down of a funding application could either be caused by the strength of local journalism within the targeted municipality, or the news outlet not matching the criteria set by the MPRT (see Appendix A for a detailed explanation of those).
Some data cleaning was required for the funding data: sometimes, funding was obtained for an area, rather than a particular municipality. Those observations were removed if too unclear which municipality was being targeted. Additionally, in some instances one news outlet would have received funding for two municipalities, or similarly two news outlets would receive one sum of money to cover one municipality. In both instances, the observations were separated in two. Following this process, of the 45 blank spots previously identified, only eight did not receive any funding, meaning that not a single outlet in Sweden has received funding to cover them. Figure 3.2 shows the amount of funding received by blank spots municipalities in Sweden. See Appendix A for a full view of all municipalities' reception of funding, including non-blank spots.

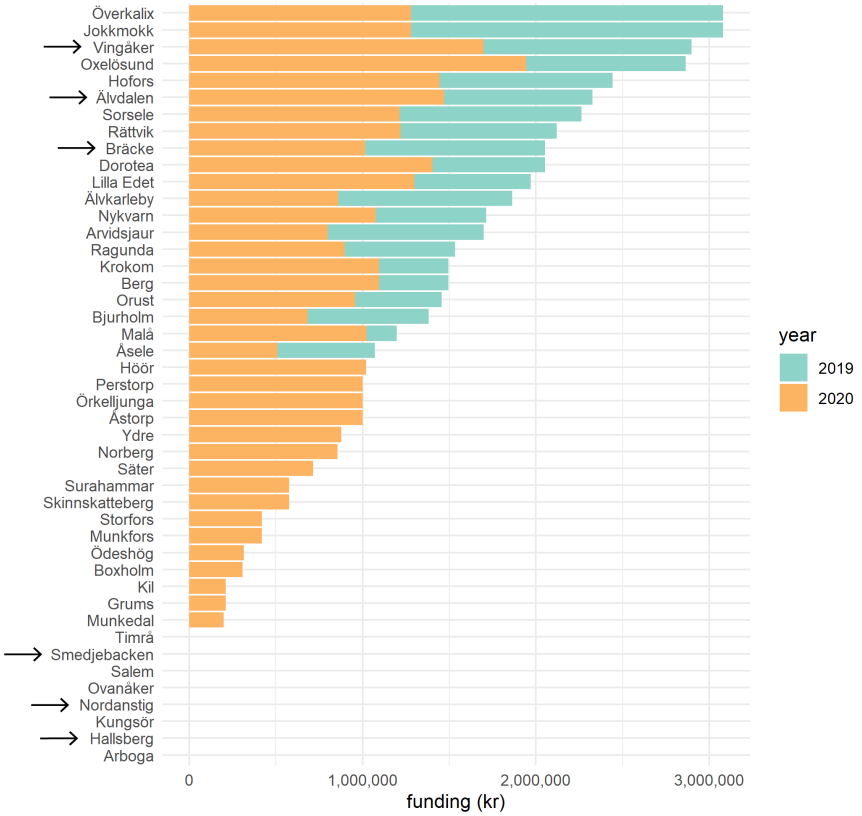


Figure 3.2: Amount of funding received by blank spot municipalities, the arrows pointing to the municipalities which are used in this study. Data from MPRT

3.1.3 Building a sample: pairing similar blank spot municipalities

Due to the time requirements of the data collection, a subset of blank spots was chosen for the scope of this study. To understand the impact of funding on the local journalism of a blank spot municipality, this project gathered a sample of articles from a total of eight blank spot municipalities (although due to challenges emerged later only six were retained for this study, as I will explain). The municipalities, four having received treatment, and four having not, were chosen through a 1:1 matching process operated on the website Kolada, run by the Swedish non-profit organisation RKA, in English known as The Council for the Promotion of Municipal Analysis. The website provides a tool to find similar municipalities based on a customizable set of over 5000 variables (RKA, 2021).

The sample was built following the principle that to observe the effect of funding on some blank spots, it would serve to have a comparative sample of non-treated blank spots, acting as a sort of counterfactual to what would have happened to the treated municipalities if they had not received the treatment (Ryan, Burgess Jr, & Dimick, 2015). This statistical approach, known as Difference-in-Differences (DID), ensures that the observable differences in the outcome variable between treated group and non-treated group are thus attributable to the main independent variable, in our case the funding. To minimise the risk of confounding effect on the differences between groups, so-called state-level confounders, one must build a control group that, based on observed data, demonstrates some ground similarity to the treated group. For this thesis, the goal was to find the most similar non-treated and treated

blank spots based on possible confounders that could have an effect on the number of articles serving a CIN, the outcome variable. Thus, literature was investigated to identify the most important municipality-level characteristics that are known to correlate with news consumption and production. Following Kekezi and Mellander (2017), I identified some individual and regional level characteristics applicable to the Swedish context. These include age, education, gender, household income, and political interest, to which I added ethnic composition and overall population, given the uneven distribution of these characteristics across Swedish municipalities (RKA, 2021).

The starting point for the construction of a sample were the eight blank spots that did not receive any funding. Due to time limitations, four of those were chosen at random to use in this study, and four were discarded. Seven, of the 5000 variables available on the Kolada website, were deemed likely to affect local journalism. Those were: population size, average age, share of women, average wage, share of foreign-born, share of turnout at the last general elections, share with a university degree or higher (see Appendix A for summary statistics). Using municipality-level data for the year 2018 (pre-treatment) on those seven variables, the most similar blank spot municipality to each of the eight blank spots not having received funding was identified.

The most similar municipality is obtained, on the Kolada website, through a similarity score, generated through a two-step process (see a description of the process in Appendix A). Following this process, four pairs were built. Table 3.1 reports the paired municipalities and their dissimilarity scores[1]. Table 3.2 shows the demographic variables for the selected municipalities.

Table 3.1: Similarity score for the selected municipalities

Control Municipality	Treatment Municipality	Similarity Score
Smedjebacken	Älvdalen	0.39
Hallsberg	Vingåker	0.44
Nordanstig	Bräcke	0.65
Ovanåker	Munkedal	0.27

Table 3.2: Demographic Variables of sampled blank spot municipalities

Municipality	Population total (n)	Share with higher education (%)	Share of women (%)	Average wage (kr)	Average age (n)	Share of foreign born (%)	Electoral turnout (%)
Älvdalen	7,121	22.85	49	1,003.800	46	8.99	86.65
Bräcke	6,376	24.67	47	823.400	47	12.27	85.04
Hallsberg	15,954	26.14	49	2,403.900	43	16.02	88.61
Munkedal	10,503	25.23	48	1,504.700	44	12.03	86.22
Nordanstig	9,517	24.42	48	1,273.500	45	9.25	83.58
Ovanåker	11,684	25.01	48	1,630.800	45	9.35	86.21
Smedjebacken	10,897	25.01	49	1,679.000	46	11.58	88.10
Vingåker	9,136	23.32	49	1,191.700	44	14.94	88.48

[1]For the municipality of Nordanstig, the most similar blank spot municipality was actually Berg, but due to the double meaning of the word in Swedish, which means mountain, the returning sample of articles in the later data collection contained a substantial amount of noise in the data, so the second most similar blank spot municipality was chosen instead (Bräcke).

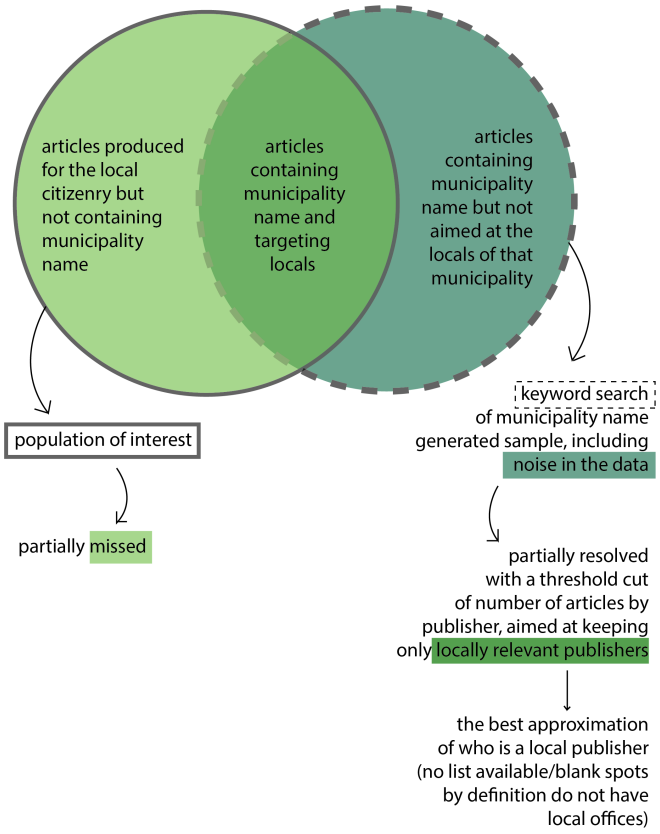


Figure 3.3: Overview of the article retrieval process and related challenges

3.2 Data

3.2.1 News Articles

Three continuous years of articles were gathered for each of the eight blank spots, for the period January 2018 to December 2020. Articles were downloaded from MedieArkivet, also known as Retriever (access obtained through Linköping University's library). The archive allows the downloading of articles (in batches of 500 articles at a time) through a keyword search function, meaning that articles featuring the input term in the title or body are displayed. As Barbera, Boydstun, Linn, McMahon, & Nagler (2021) point out, it is important to understand the limitations a particular keyword presents in relation to the relevance and representativeness of the population of interest. In this research, the articles were collected through a keyword search of the municipality name, in a similar fashion to prior research also attempting to map journalism produced within a municipality (Nygren & Schjærff Engelbrecht, 2018). Filter criteria in the interface allowed to limit article selection to those being published in the press or in online format, and belonging to a regional or municipal press. Thus news stories published on TV, radio, magazines or major national newspapers were excluded from the search. Does the resulting corpus overlap with the population of interest? Not exactly. The population of interest, in this research, is the general local journalism coverage for a municipality. By selecting articles including the name of the municipality in the text, two limitations are identified. First, articles relevant to a municipality that do not feature its name in the text are missed entirely. Second, articles mentioning the municipality, yet not targeting that specific municipality, are included. See Figure 3.3 for an overview.

The data collection generated a corpus of approximately 80000 articles, stored in text file format. The data was converted to csv, cleaned and restructured to a dataframe featuring one article per row, with information about the publisher, the publication date, and the municipality reported. A cross-check analysis revealed that for two funded publishers, Siljan News (covering Älvdalen) and Tidningen Västsverige-Västbygden (covering Munkedal), Retriever did not capture any article. This is due to the fact that Retriever does not track all news outlets in the country, despite being the most comprehensive digital archive for news articles in Scandinavia (Retriever, 2021). While Älvdalen had more than one outlet having received funding to cover it, Munkedal solely relied on Tidningen Västsverige-Västbygden for funding, and thus I discarded the pair Ovanåker-Munkedal from the analysis. Including them would have invalidated the analysis on the impact of funding on local journalism at the municipality level, given that articles from the outlet having received the funds are missing from the article dataset (unless one believed that journalism would change across all outlets, regardless of which outlet receives the funding). This reduced the corpus size to 64472 articles.

While there was no identified way to solve the issue of missing some of population of interest within the time limitations of this thesis (e.g. by testing several keywords, which would have increased the data collection time requirements), some measures were taken in regards to the noise in the data. The challenge here is to attempt to remove the largest possible number of articles that included the municipality name but were not written for that specific local citizenry. The easiest way to approach the problem would be to filter out all articles published by a newspaper targeting a different municipality. However, as mentioned previously, there is no record in Sweden as to which municipality a news outlet targets. Being the sampled municipalities blank spots, none of the municipalities for which articles were gathered had a single editorial office within them (at least during the pre-treatment period), making the task of identifying relevant newspapers even more challenging. Thus, the following approach was used. A count of articles by publisher for each municipality revealed that the data is constituted by just a few publishers who produce most of the articles featuring the municipality name, while most publishers, in the span of three years, mentioned the municipality only a handful of times. Many of those rare publishing titles were indeed papers covering completely different geographical areas. Thus I chose a threshold number n to make a cut: if a publisher had only mentioned a municipality less than n times in the three year period for which articles were gathered, then those articles (and publishers) were discarded from the analysis. A compromise between loss of publishers and loss of articles guided the choice in the threshold value, as well as a look at the number of articles published by the funded publishers. The threshold value was set to $n = 40$. Figure 3.4 gives an overview of the outcome of the cutoff: a large number of publishers has been discarded, in favour of a smaller number of frequent publishers.

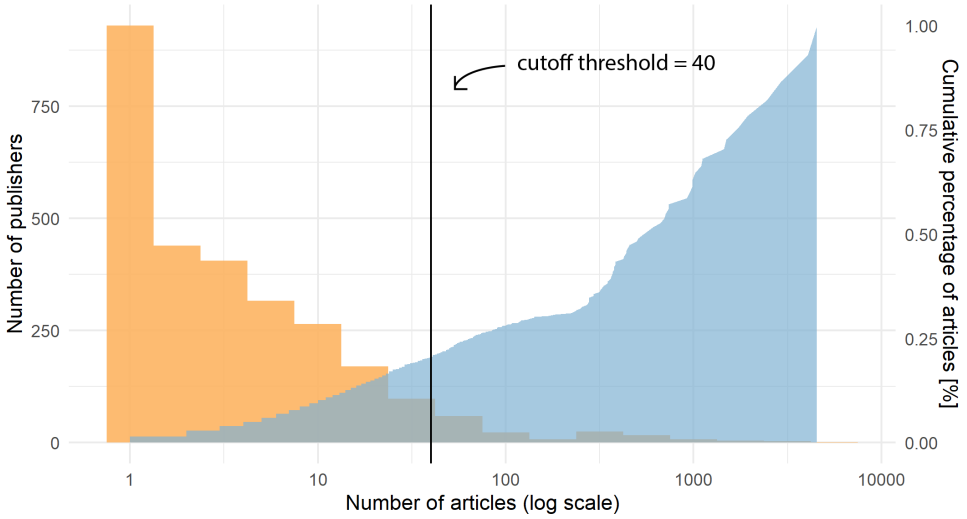


Figure 3.4: Number of articles by publisher, against cumulative distribution of articles. Data from Retriever

After the cutoff was applied, the total number of articles remaining in the corpus was 51209. Table 3.3 shows how the articles are distributed across municipalities, in pre-post treatment period, and by funding status. Information about which publishers received funding and when can be found in Appendix A.

Table 3.3: Number of articles by municipality and publishers, divided by funding allocation (yes = 1, no = 0), and treatment period (pre = before 1st July 2019, post = after 1st July 2019)

Municipality	Funding	Pre	Post	Publishers	Total
Älvdalen	1	29	39	1	68
Bräcke	1	2527	856	3	3383
Vingåker	1	4559	3553	2	8112
Älvdalen	0	4828	3832	18	8660
Bräcke	0	2095	1545	17	3640
Vingåker	0	1279	1284	18	2563
Hallsberg	0	5489	4502	26	9991
Nordanstig	0	2867	2216	10	5083
Smedjebacken	0	5200	4509	18	9709

3.3 Method

3.3.1 Supervised Machine Learning

The task of understanding the impact of funding on local journalism' supply of CINs requires the obtainment of a quantitative measure capturing whether an article serves or not a critical information need. Prior attempts in the CINs literature have used manual annotation (Napoli, Weber, MCcollough, & Wang, 2018), or availed of the insights from the manual annotation to build a dictionary of keywords relating to each CIN. The researchers then used that

to classify whether an article in the TodayIn news feature of Facebook covered a critical information need (Weber, Andringa, & Napoli, 2019). Due to the lack of prior research on CINs in Sweden, it was not possible to build or use a similar approach. Even if Weber, Andringa, and Napoli (2019)'s dictionary was available, differences in contexts between US communities and Swedish communities would have likely made the dictionary a non-accurate identifier of the keywords associated to each CIN in the sampled municipalities of this study. For this reason, in this thesis I instead avail of supervised machine learning (SML) to train a model able to predict if an article fulfills a critical information need based on the semantic similarity to a hand-annotated set of articles from the same corpus, and tag it with a label specifying which particular CIN is covered. While untested in the research on blank spots in Sweden, SML has been adopted elsewhere in journalism studies as a tool to facilitate the classification of text based on its semantic features (Burggraaff & Trilling, 2020).

For the model to learn how to classify articles, it requires to be presented with some cases that are representative of the data it will later predict on, and that are associated with the correct labelling answer. For this reason, a random sample of 1000 articles was extracted from the full dataset and hand-coded. For each article, one label was applied describing the article classification category. Eleven labels were chosen: eight, one for each CIN, and an additional three, one for articles about sport, one for obituaries, and for articles that did not fit any particular category. The addition of the three extra categories is based on an evaluation of the type of articles not fitting any CIN, and it is analogous to the procedure followed by Weber, Andringa, and Napoli (2019). For each article only one label was chosen, meaning that whenever an article would suit more than one label, I annotated based on which label seemed more prominent within the article.

There are several choices one can make in relation to supervised machine learning model. Particularly, one can choose what parameters to use to train the model, for example word vectors or word frequencies, and what type of model to use, for example a logistic regression, a naive-bayes probabilistic model, a support-vector machine, or a decision tree algorithm. In this thesis, the model was built using the fastText library for R [2]. FastText is an open-source library for Natural Language Processing (NLP) developed by the Facebook AI Research lab in 2016, that tackles the tasks of text classification and learning word vector representations (Joulin, Grave, Bojanowski, & Mikolov, 2016). The text classification model has become highly popular thanks to reduced computational requirements while maintaining high performance, comparable to the one of neural networks, which in contrast tend to be rather slow (Joulin, Grave, Bojanowski, & Mikolov, 2016). Its implementation is simple and flexible, allowing the researcher to tune the model to a preferred set of parameters, relating both to the parameters that are used to train the model, and to the hyperparameters of the model itself.

FastText uses a bag of n-grams as features, where n-grams are the concatenation of *n* consecutive words in a given text. This is generally preferred to the well-established Bag of Words (BOW) approach, as n-grams preserve the semantic information which emerges by analysing words in relation to their neighbors in a sentence. The mapping of the n-grams is done using the *hashing trick*, which decreases the memory requirements and increases speed efficiency

[2]Different models were attempted, namely a series of models, including random forest, multinomial logistic regression, and support vector machines, that used the TF-IDF of the annotated set to form a dictionary of keywords for each category. Using the open source FastText library of pre-trained word embeddings for the Swedish language, the models calculated Concept Mover´s Distance (CMD) between each keyword and each document, by averaging the distance between each word in a document and a keyword in the embedding space. The CMD yielded slightly worse results than the FastText model, and the risk of overfitting due to having selected the TF-IDF words based on the test set as well as the train set - which I did not manage to avoid having to do - led me to select FastText in the end. See Stoltz and Taylor (2019) for an explanation of CMD, and Appendix A for an overview of the process. The code for the additional models can be found in Appendix B.

(Joulin, Grave, Bojanowski, & Mikolov, 2016; Weinberger, Dasgupta, Langford, Smola, & Attenberg, 2009). From a technical perspective, the model translates each n-gram of preferred length, and each label, to a vector representation that describes the position of those text elements in a hyperdimensional space. The n-gram embeddings constituting a document are then averaged to return a document vector. The model is thereafter attempting to reduce the distance between the vectors of each label and its associated document. Each document vector is input into a softmax function f with each of the possible labels, generating a series of scores describing the probability that the text belongs to each of the labels respectively. The softmax function is used to minimise the negative log-likelihood of the labels, as presented by the fastText authors Joulin, Grave, Bojanowski, and Mikolov (2016) in their paper, with the log-likelihood being:

$$-\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n log(f(BAx_n)) \tag{3.1}$$

Where
x_n = the normalised bag of features for the nth document
y_n = the label for the nth document
f = the softmax function
A and B = the weight matrices, A being the matrix retrieving the word embedding, and B being the linear output transformation

Using stochastic gradient descent, combined with a linearly decreasing learning rate, the vector coordinates of a document and its label are brought as close as possible to each other, maximising the probability score for the correct label. A simple overview of the algorithm structure is presented in Figure 3.5.

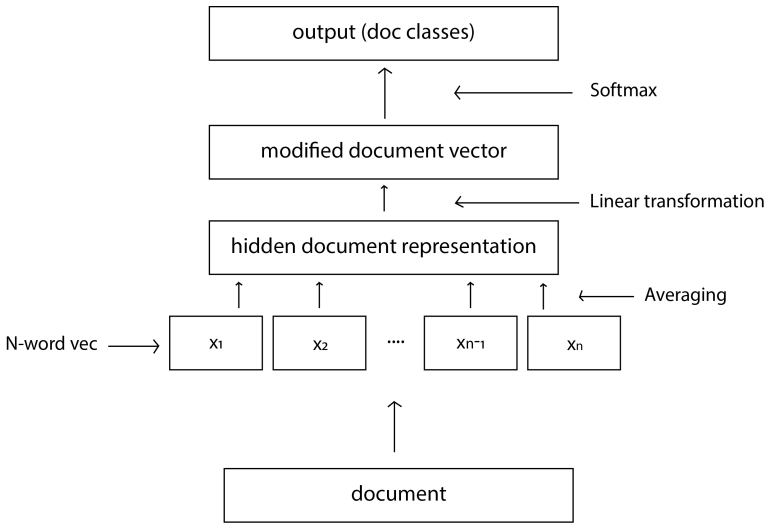


Figure 3.5: The structure of the fastText supervised algorithm for text classification. Following Joulin et al. (2016)

The model was configured to the following hyperparameters settings. The *loss* function set to *softmax*, after testing it against and yielding better results than its counterpart *hs*, hierarchical softmax, that is considered to be particularly efficient at handling class imbalances in multi-label classification tasks. The *learning rate*, which controls the amount at which the weights

yielding the highest reduction in error at each iteration should change, was set to 0.5. Faster learning rates, close to 1, can lead to poor performances on new data, while smaller learning rates, close to 0, can require a higher number of iterations given the small learning rate. The number of iterations, also known as epochs, was set to 7000, after testing a range of values. The word vectors length to be generated was set to 50, indicating the number of dimensions constituting the ngram vector representation. The maximum length of ngrams used was set to 10. The number of buckets was set to 10000000, where a bucket refers to the size of a vector storing the hashes of a an n-gram. The remainder parameters were left as default, and a full specification of those can be found in Appendix A. Finally, the 1000 annotated articles were randomly split into a training set constituting 80 percent of the data, and a test set of the remainder 200 articles. The model was trained with the train set, and evaluated on the test set, before being applied to the full dataset for classification.

3.3.2 Statistical Analysis

The analytical strategy adopted in this study is built around the extensive time requirement of the data collection process, constraining the number of municipalities possible to analyse. The choice of six blank spot municipalities was guided by the principle that to make causal inferences about the impact of funding on local journalism one would need to compare different groups, one having received the treatment and one not (Rubin, 1974). As observable in Table 3.3, there is a substantial imbalance in the number of articles published in Älvdalen by its funding receiving publisher (n = 39), and the remainder ($n \geqslant 8000$). For this reason, the pair Älvdalen-Smedjebacken was excluded from the statistical analysis.

To test both hypotheses I adopt a statistical strategy known as Difference-in-Differences (DID). DID is an attempt to mitigate the risk of observed and unobserved confounders that arise in the absence of a randomized controlled trial. In those, subjects are randomly allocated treatment, and differences between a treatment group and control group are used to capture the magnitude of the causal effect of the treatment on the outcome of interest. When treatment is not randomly allocated, the estimate of the causal effect could easily be confounded by unobserved factors related to the subject's characteristics or its environment (Strumpf, Harper, Kaufman, & Oakes, 2017).

DID is a quasi-experimental design that captures, as the estimate of the causal effect, the difference in the change in the outcome variable post and prior to the treatment implementation, between a group that has received the treatment and one that has not. In other words, what happened to the control group is used as a counterfactual of what would happened to the treatment group, may it had never received it, while still accounting for original differences between the two groups. DID is considered a valid design when the parallel-trend assumption (PTA) is fulfilled, meaning that the statistical difference of the outcome variable across the treatment and control units in the pre-intervention time period is insignificant (Callaway & Sant'Anna, 2020). The causal effect is estimated through a linear regression of the form:

$$Y_{ijt} = \beta_0 + \beta_1 Treat_j + \beta_2 Post_t + \beta_3 Treat_j Post_t + \varepsilon ijt \tag{3.2}$$

Where
Y_{ijt} is the outcome for subject (publisher/municipality) i in group j (the treatment allocation group it belongs to) at time t (a specific year quarter, in my case)
$Treat_j$ is a dummy identifying the group exposed to the treatment
$Post_t$ is a dummy variable identifying the first time funding was distributed

$Treat_j$ x $Post_t$ is the interaction between the two
ε_{ijt} is the error term

At the heart of the estimation is the interaction term between Treat and Post, which equals the Average Treatment Effect on the Treated (ATT). In its simplest form, it is the difference across the two time periods in the expected value of y for a particular group, minus the difference across the two time periods in the expected value of y given belonging to the other group. In this thesis I use a variation of the DID design that uses group-time ATT to account for time and group fixed effects, which introduces the ability to account for multiple time periods, and treatment allocations, following Callaway and Sant'Anna (2020):

$$ATT(g,t) = E[Y_t(g)Y_t(0)|G = g] \tag{3.3}$$

The advantage of this setup within this thesis is that it introduces time fixed effects, enabling to see the effect of treatment at multiple time points, and thus, compared to the canonical DID design of two time periods, it controls for time trends that might happen within the pre and post intervention period. Here, I use yearly quarters as time points, and individual ATT estimates are obtained for the treatment group for each quarter. Those can then be aggregated in one averaged ATT effect for the entire post-intervention period, to obtain a general estimate of the effect of the treatment overtime (Callaway & Sant'Anna, 2020):

$$\theta_W^O = \frac{1}{\kappa} \sum_{g \in G} \sum_{t=2}^{\tau} 1\{t \geqslant g\} ATT(g,t) P(G = g | G \leqslant T) \tag{3.4}$$

where in this formula Callaway and Sant'Anna (2020) apply weights to group sizes to account for different groups receiving treatment in different periods, but, as I will show shortly, in my case I only have one treatment group, having received treatment at only one point in time.

I run two sets of analyses, one at the publisher level and one at the municipality level. The publisher-level analysis should be interpreted as the effect of funding on the specific outlets having received it, while the municipality-level analysis offers instead the opportunity to make a statement on whether the overall state of journalism for a blank sport municipality has changed. In the publisher-level analysis I use as treatment group the funding-receiving publishers in Bräcke and Vingåker, and the control group is the untreated publishers in the same municipalities, thus I exclude Hallsberg/Nordanstig from the analysis. For the municipality-level analysis, the treatment groups includes all the publishers writing within a municipality in which at least one publisher has received treatment (those are Bräcke and Vingåker), and the control group is comprised of all the publishers within the untreated municipalities of Hallsberg and Nordanstig.

For each level of analysis, I run the same three models: the first one aims at testing H1, that *the allocation of funding will have increased the volume of critical information needs articles in those municipalities within which some publishers have obtained funding, and particularly for those funding-receiving publishers*. It does so by testing the impact of funding on the number of CINs articles. In the second model, I include all articles, not just the CINs ones, to assess the robustness of my results when I take into account the entire journalistic production. In the third model, I test for H2, that *the allocation of funding will have increased the overall amount of articles about Civic Information and Politics*, by measuring the impact of funding on the number of articles about those categories.

To fit the DID design, the data was transformed to take on a panel data structure, where for each publisher an aggregated measure of the number of CINs articles within each yearly

quarter was obtained. For each publisher there are thus twelve time points, half which count as pre-intervention and half as post. The *dependent variable* is the count of the number of articles (different for each of the three models I run) produced by a given publisher within a yearly quarter. *Treat* is a dummy identifying the group exposed to the treatment (publisher or municipality, depending on the model). *Post* is a dummy variable identifying the first time funding was distributed, which was indicated as the 1st July 2019, a conventional date chosen as the MPRT report simply states funding was distributed during the summer of 2019 (MPRT, 2019). Because nearly all funded publishers received funding in 2019 and in 2020, the DID model simply considers the treatment time as one continuous period beginning summer 2019, rather than separating the two funding allocations. However, for two publishers funding was only allocated in one year. Sundsvalls Tidning received funding in 2019, and in this setup it was considered to remain treated during the second half of 2020, due to not knowing when precisely the funds were used and what for specifically. On the other hand, Östersunds-Posten received funds in the second part of 2020, and I chose to exclude him from the analysis, as for all remainder publishers the effect started in the third quarter of 2019. While the DID setup used in this analysis would have allowed for having different groups treated at different points in time, having only one publisher in the second treatment period group did not constitute a sufficiently large sample size. Finally, for all those publishers which had zero articles published within a specific year quarter, those appeared as missing time points in the data. I thus introduced those missing time points for the specific publishers, where I inserted 0 as the total count of articles published in that particular time period. Table 3.4 gives an overview of the sample sizes for each of the three models, at the two levels of analysis.

Table 3.4: Number of observations in the three models, at the two level of analysis, in brackets the number of publishers

Dependent Variable	Publisher	Municipality
CIN	468 (39)	888 (74)
All articles	468 (39)	888 (74)
Politics and Civic Information	468 (39)	876 (73)

As we see from Table 3.4, I lose one publisher at the municipality level, when the dependent variable is only articles about politics and civic information, due to the publisher not having written any articles, across all twelve time periods, for those two article categories. In the Appendix I include results for the other two models where I exclude this publisher (Göteborgs-Posten, within Hallsberg), to assess the consistency in results when I account for the difference in sample sizes.

4 Results

This section introduces the results obtained at various steps of the project. First, the SML model is evaluated, and the outcome of the classification of the dataset is presented. Thereafter, results from the difference-in-differences analyses are reported.

4.1 Model Evaluation

The following metrics describe the model performance on the test set. The model accuracy was 0.665. Kappa was 0.624, indicating the predictive accuracy of the model adjusted for the possibility to label correctly by chance (Nwanganga & Chapple, 2020). Precision and recall were 0.665 each. The former describes the positive predictive value, the proportion of correctly predicted values of a given label against the total number of predictions for that same label and is calculated as follows (Nwanganga & Chapple, 2020):

$$precision = \frac{TP}{TP + FP} \tag{4.1}$$

Instead, recall measures the correct number of predictions for a given label against the total number of false predictions and true positives made for that label (Nwanganga & Chapple, 2020):

$$recall = \frac{TP}{TP + FN} \tag{4.2}$$

The F1 score was 0.678, and it describes the harmonic mean of precision and recall, each of them given equal weight. The score provides a handy way to boil down a model's performance into one single number, proving particularly useful when evaluating different models (Nwanganga & Chapple, 2020):

$$F1score = \frac{2\ x\ precision\ x\ recall}{precision + recall} \tag{4.3}$$

Other two metrics, sensitivity and specificity, are reported. Both ranging between 0 and 1, they offer a tradeoff: models that have high sensitivity give more emphasis to the share of correct classifications, while focusing on improving the specificity is desirable when one does not want to misclassify a high share of labels (Nwanganga & Chapple, 2020). Sensitivity

was 0.665 (it is the same as recall) and specificity was 0.966, referring to the correct share of negative classifications, measured as follows:

$$specificity = \frac{TN}{TN + FP} \tag{4.4}$$

Finally, Figure 4.1 presents an overview of the classification of the model on the test set.

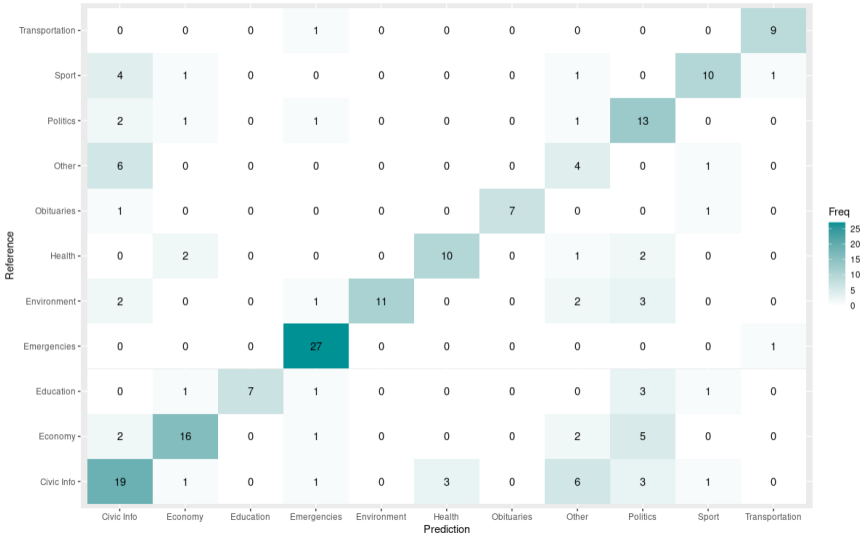


Figure 4.1: Confusion matrix of the test set

Given the presence of eleven categories, a random classifier would have a 0.09 accuracy score, which, compared to our accuracy metric of 0.665, indicates that the model does an acceptable job at classifying the articles. Some categories are more easily predicted than others, for example Emergencies and Transportation. Other categories, such as Civic Information and Other, get more easily misclassified. These two in particular got misclassified among each other, which was somewhat expected, as both categories contained a varied sample of stories, the difference being how critical the information in the article was, which the algorithm is not directly trained to detect. If we just look it from an aggregate perspective, with the critical information needs articles on the one side and the remainder categories in the other, what we see is that a few CINs articles are misclassified as belonging to the Other category, and marginally in the Sport category as well, but none was misclassified as belonging to the Obituaries category. On the other hand, some non-CINs articles were classified as a CIN category, mostly for the Other and Sport category.

4.2 Text Classification

The SML model was used to predict the category of the articles in the full dataset. Then, by aggregating the CIN categories together, I was able to obtain a picture of the relevance of articles in the sampled municipalities. Figure 4.2 shows the mean number of articles across municipalities, divided by whether they fit a CIN or not (see Appendix A for a table of descriptive statistics). The grey shaded areas represent the confidence intervals. The plot shows that overtime the number of CIN articles is consistently larger than the number of non-CIN

stories. It also shows that there seemed to be a decrease in the number of CIN articles in the first half of the observed time period.

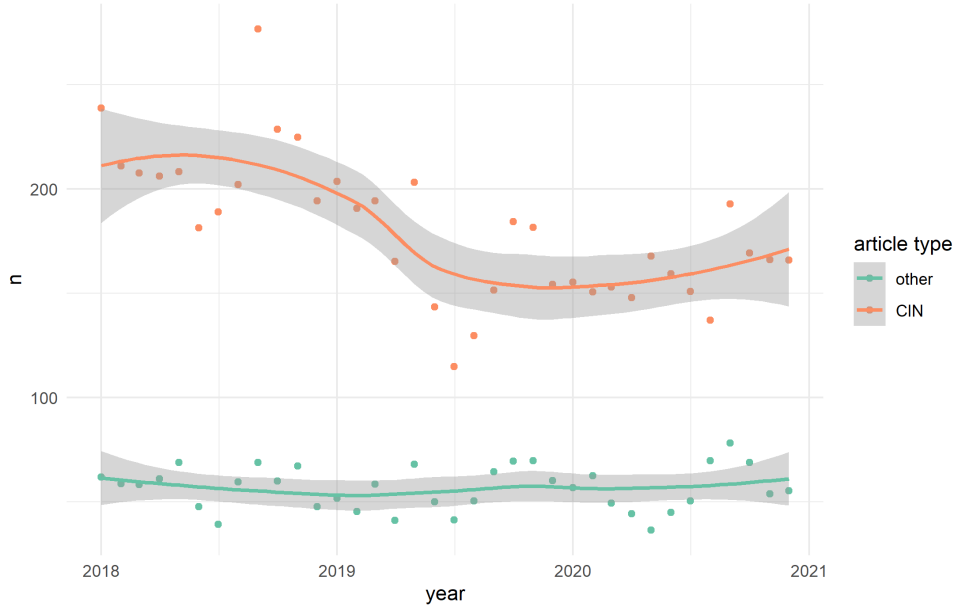


Figure 4.2: Mean number of CIN and non-CIN articles overtime, each dot the monthly mean across municipalities

Figure 4.3 instead looks at the individual labels, and shows that some labels are more predominant than others (e.g. civic information and politics), while others occupy less journalistic space (e.g. obituaries and transportation). The Figure also shows that the share of a particular label is quite similar for different municipalities, with the biggest variation in share found in the labels of politics, sport, environment and obituaries.

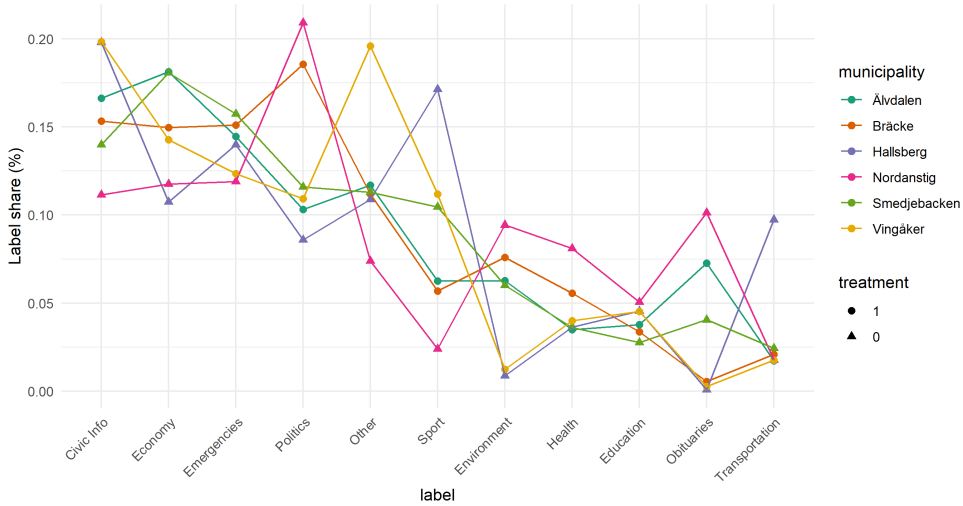


Figure 4.3: Share of predicted labels on the full dataset (each point a municipality, distinguished between treated and untreated)

As Figure 4.4 shows, the pattern in terms of label share does not vary substantially by municipality, showing consistency in the classification across publishers covering different blank spots. Some of the observed differences are attributable to inherent characteristics of the municipality. For example, the municipality of Hallsberg, which has an important train station connecting Stockholm and Gothenburg, as well the north and south of Sweden, has also the highest share of articles about transportation. Finally, Figure 4.4 shows that most articles in the dataset fulfill a critical information need, consistently across municipalities, while the remainder categories only capture around a fourth of all the articles.

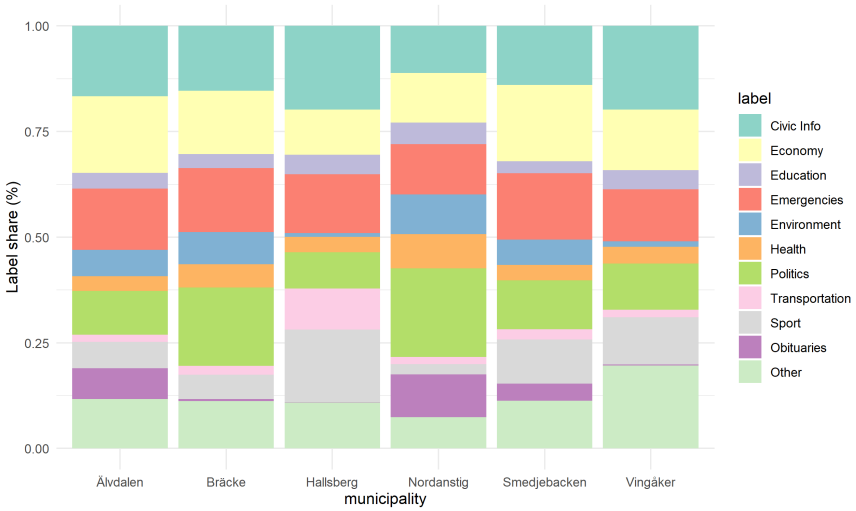


Figure 4.4: Predicted label shares by municipality, ordered by CIN labels, followed by the additional label. Älvdalen, Bräcke and Vingåker being the funded municipalities; Hallsberg, Smedjebacken and Nordanstig the unfunded ones

When looking at the share of labels within different publishers, we see a similar picture. Some publishers have a high share of one label, sometimes nearly as high as 1.0, indicating that different publishers within the dataset provide differing levels of variation across coverage, perhaps due to being a specialised publication (e.g. a sports magazine). However, for each label the median publisher share of that one category is well below 0.25 (see Figure 4.5). The Figure also shows that for all labels, up until the third quartile of the data is below 0.25. A guess would be that for the publishers showing high shares of one particular label, there would probably not too many articles covering that particular municipality, mostly because they likely do not constitute the most important paper for that blank spot. Nonetheless, because of the threshold cutoff of 40 articles operated in the pre-data collection, even a publisher which produced only articles about sport, would have had to at least produce 40 articles about that municipality and have them labelled as sport.

In this section, I have presented an overview of the picture of local journalism and critical information needs in the selected blank spots. The takeaway from the classification is that blank spots are indeed provided with critical information needs stories, consistently across municipalities, by means of local newspapers, and that indeed those type of stories account for the majority of local newspapers' production.

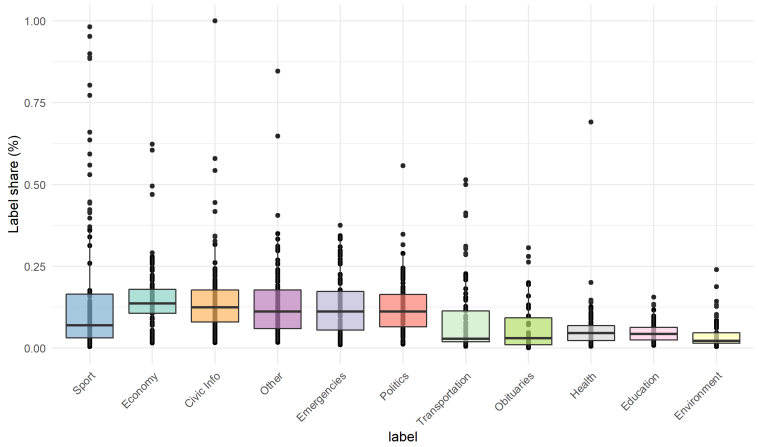


Figure 4.5: Share of a predicted label for each publisher (each publisher a dot)

4.3 Statistical Analysis

In this section I present the results from the measurement of the impact of funding on the number of articles about CINs and about civic information and politics. An exploratory assessment of the data is done in Table 4.1, which shows that there appears to be a mixed picture across municipalities in relation to the difference in the number of articles before and after funding allocation. For all treated publishers we see a decrease in the overall number of articles published post intervention, regardless of whether they fulfill a CIN or not. Between the unfunded publishers, a decreasing pattern is generally predominant, with the exception of the publishers in Nordanstig, where no outlets have received funding, and the unfunded publishers in Vingåker, which both see an overall increase in the number of stories that do not fulfill a CIN.

Table 4.1: Overview of the changes across CIN (category = 1) and non-CIN (category = 0) articles before (Pre) and after (Post) funding allocation

Funding	Municipality	Category	Pre	Post	Difference	Proportional Difference
0	Bräcke	0	261	220	-41	-15.7
0	Bräcke	1	1834	1325	-509	-27.8
0	Hallsberg	0	1360	1254	-106	-7.8
0	Hallsberg	1	4129	3248	-881	-21.3
0	Nordanstig	0	446	515	69	15.4
0	Nordanstig	1	2421	1701	-720	-29.7
0	Vingåker	0	355	431	76	21.4
0	Vingåker	1	924	853	-71	-7.7
1	Bräcke	0	366	173	-193	-52.7
1	Bräcke	1	2161	683	-1478	-68.4
1	Vingåker	0	1289	1084	-205	-15.9
1	Vingåker	1	3270	2469	-801	-24.5

When it comes to number of articles, journalism is highly dependent not just on financial resources but on the stories that are newsworthy, the exceptional events that redirect reporters' attention, and that perhaps require several follow-up articles. Overtime, rates of news are

thus susceptible to increases or decreases. For example, we see a peak of political stories around the time of the General Elections of 2018, and similarly we see an increase in health stories following the beginning of the Covid-19 pandemic (see Figure 4.6). The pre-treatment period was characterised by a high number of environment and emergencies news, most likely due to the 2018 Sweden wildfires, which took place as a result of particular hot and dry summer (Krisinformation, 2018). Overall, this also provides some validity in the coding scheme, by showing consistency between external events and the predictions.

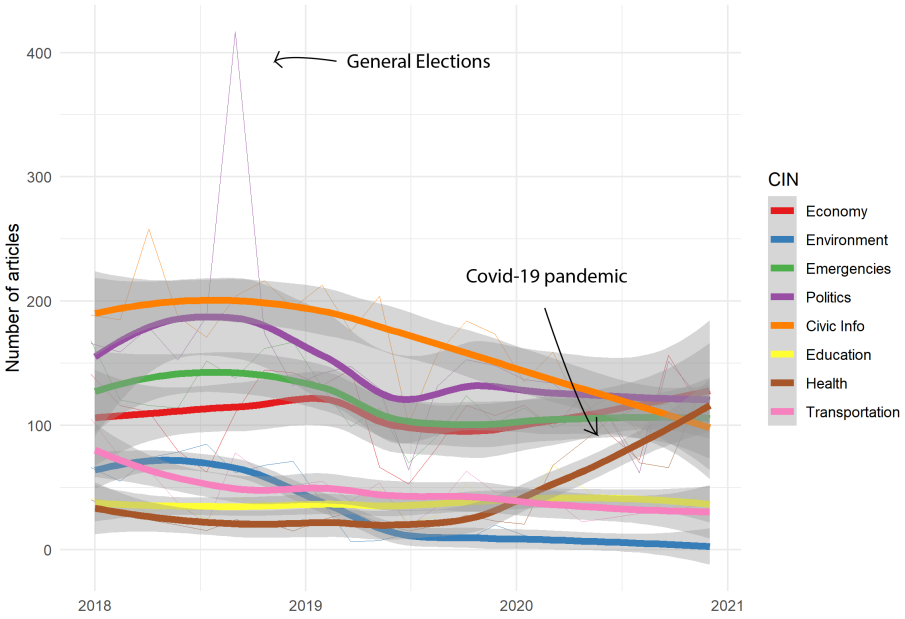


Figure 4.6: Time trend of individual CINs in the article dataset. In bold the smoothed conditional means, and the thin lines the number of articles for a particular CIN category at a given point in time

4.3.1 Publisher-Level Analysis

In this section, I present the results from three models aimed at providing an overview of the impact of funding on various dimensions of local journalism for the treated publishers, using the untreated publishers within their same municipality as the control group.

In the first model, I test the impact of funding on the number of articles about a CIN. The aim is to test H1, that the funding has increased the volume of articles about a CIN, for those publishers that have received funding. Figure 4.7 shows that the funding has had no impact on the number of CIN articles at the publisher-level, which is shown by the blue confidence bands crossing over the zero in the plot. The red confidence bands crossing the zero instead provide an understanding that the mean difference in the outcome variable between the two groups at each time period in the pre-intervention phase does not significantly change in respect to the first time period (2018.1), and can be understood as a sign that the parallel-trend assumption holds in respect to the initial differences between the two groups. Overall, however, we also see that, although insignificant, the direction of the coefficient seemed to be trending positively, and the confidence bands of the coefficient become more steady post-intervention.

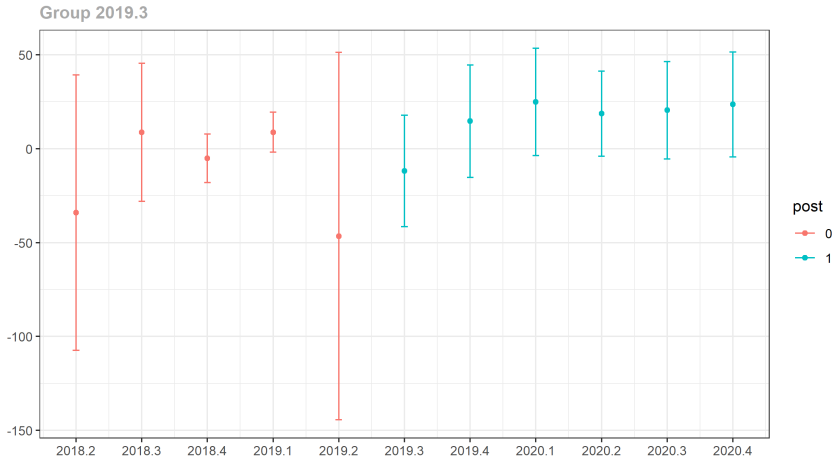


Figure 4.7: ATT coefficients showing the impact of funding at each year quarter on the number of CIN articles, for treated publishers against untreated publishers within the same municipality

In the second model, I test the impact of funding on the cumulative number of articles for a given publisher in a specific year quarter, regardless of category. The aim is to provide a robustness check of my previous model, and see whether when we account for the entire journalistic production, the picture changes significantly. Figure 4.8 shows a positive trend overtime, yet insignificant. The parallel-trend assumption seems to hold.

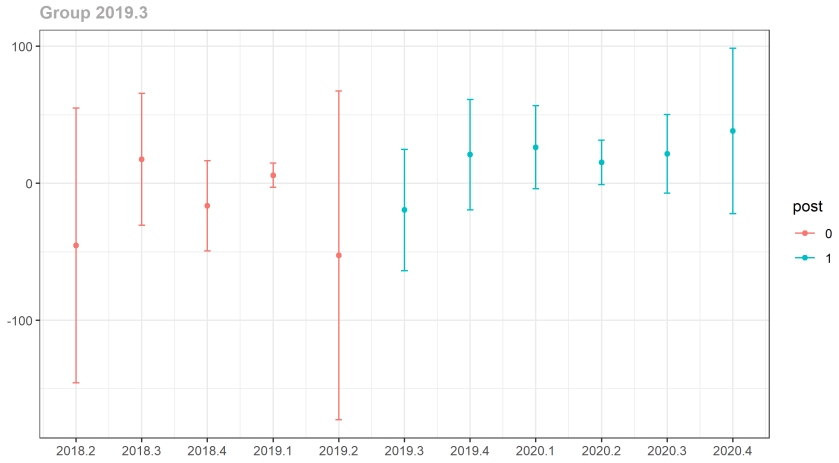


Figure 4.8: ATT coefficients showing the impact of funding at each year quarter on the number of articles overall, for treated publishers against untreated publishers within the same municipality

Lastly, in the third model I test the impact of funding on the average number of articles about Politics and Civic Information published by a treated outlet. The aim is to test H2, that the funding has increased the volume of the civically relevant categories of Civic Information and Politics. Figure 4.9 shows an insignificant impact of funding on the volume of civic and political journalism for the treated publishers, and the parallel-trend assumption seems to

hold. Compared to the two models above, we do not see the same positive trend occurring, instead the post-intervention coefficients remain close to zero.

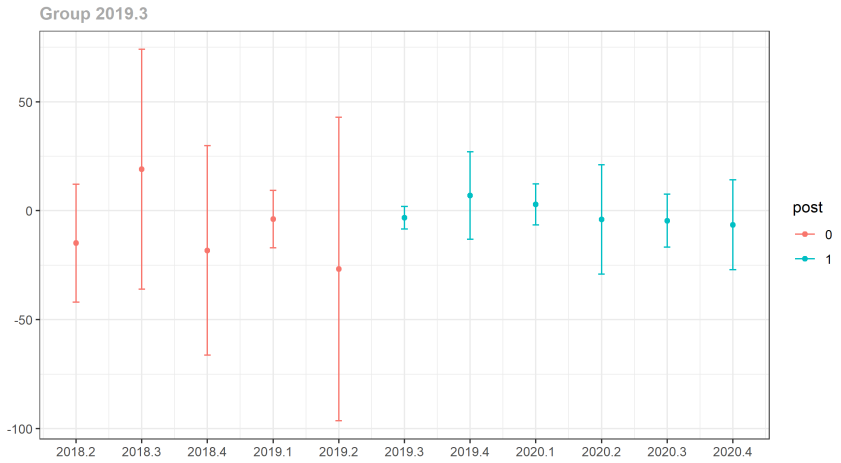


Figure 4.9: ATT coefficients showing the impact of funding at each year quarter on the number of articles about civic information and politics, for treated publishers against untreated publishers within the same municipality

Finally, Table 4.2 shows the averaged ATT coefficients for the overall intervention period. See Appendix A for a detailed overview of each model output, including the specific coefficients at each year quarter.

4.3.2 Municipality-Level Analysis

In this section, I instead run the same three models but at the municipality-level, the aim to see whether the funding has generated change, in terms of the three tested dependent variables, on the municipality in which some of its publishers have received funding. The treatment group is all publishers within a municipality where some outlets have received funding (that is, all publishers in Bräcke and Vingåker), and the control groups is the publishers in the municipality of Hallsberg and Nordanstig.

In the first model, I test the impact of funding on the number of articles about a CIN, again to test H1. Figure 4.10 shows that by yearly quarter the funding has had no impact on the number of CIN articles and that the parallel-trend assumption holds. Like in the publisher-level analysis, the coefficients, although insignificant, head in a positive direction, and the confidence intervals become more steady post-treatment.

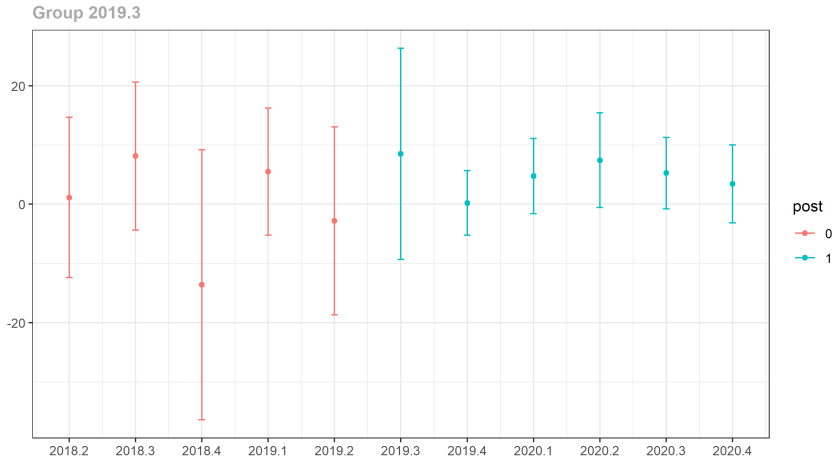


Figure 4.10: ATT coefficients showing the impact of funding at each year quarter on the number of CIN articles, for all publishers in Bräcke and Vingåker against all publishers in Hallsberg and Nordanstig

In the second model, I test the impact of funding on the cumulative number of articles, regardless of category. Figure 4.11 shows, similar to above, an insignificant impact of funding on the overall volume of journalism for the treated municipalities, and the parallel-trend assumption seems to hold.

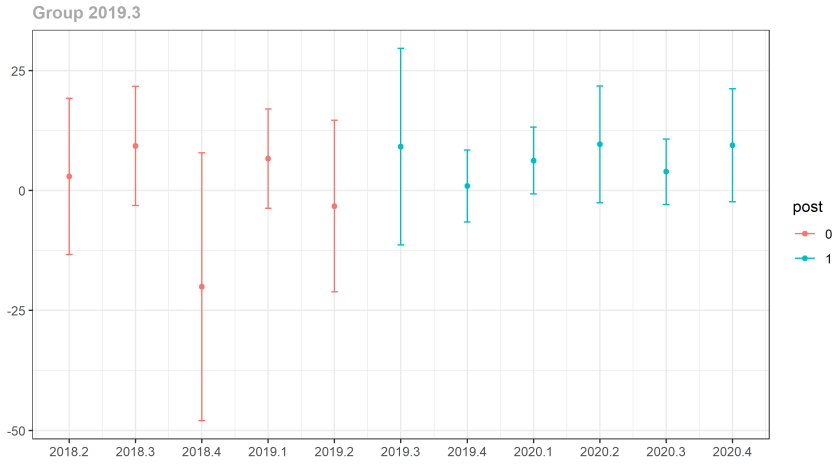


Figure 4.11: ATT coefficients showing the impact of funding at each year quarter on the number of articles overall, for all publishers in Bräcke and Vingåker against all publishers in Hallsberg and Nordanstig

Lastly, in the third model I test the impact of funding on the average number of articles about Politics and Critical Information, the aim being testing H2. Figure 4.12 shows an insignificant impact of funding on the volume of political and civic information articles for the treated municipalities, and the parallel-trend assumption seems to hold.

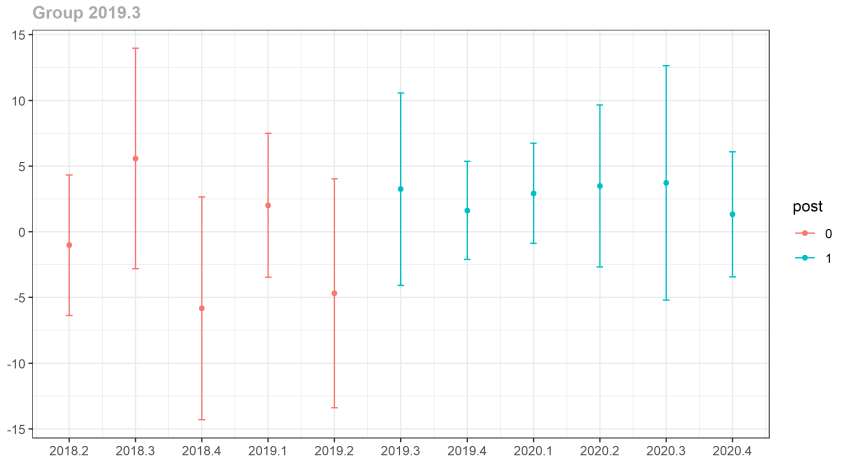


Figure 4.12: ATT coefficients showing the impact of funding at each year quarter on the number of articles about civic information and politics, for all publishers in Bräcke and Vingåker against all publishers in Hallsberg and Nordanstig

Table 4.2: Averaged ATT coefficients for the overall intervention time period. Robust standard errors are reported

Publisher-Level			
dependent var	ATT	Std. Error	[95% Conf. Int.]
CINs	15.07	8.82	[-2.23, 32.36]
All articles	17.10	10.85	[-4.16, 38.36]
Civic Info and Politics	-1.42	4.10	[-9.45, 6.61]
Municipality-level			
dependent var	ATT	Std. Error	[95% Conf. Int.]
CINs	4.92	2.39	[0.23, 9.61] *
All articles	6.55	2.85	[0.96, 12.14] *
Civic Info and Politics	2.72	2.09	[-1.37, 6.81]

Signif. codes: '*'
Control Group: Never Treated
Est. Method: Doubly Robust

Finally, Table 4.2 shows the averaged ATT coefficients for the overall intervention period. Like above, see Appendix A for a detailed overview of each model output, including the specific coefficients at each year quarter.

The conclusion to be drawn from the above analyses is that I find no support to H1 and H2, at the publisher level. Thus, I here interpret the funding as having had no significant impact in the volume of journalistic content of the treated publishers when compared to the untreated publishers within the same municipality. Insignificance is found both when looking at overall journalistic production, or considering only CINs articles, but also when selecting the two categories that are deemed most associated to local newspapers and the watchdog role of the local reporter: civic information and politics. This holds both for the individual time period ATTs, and in the joint estimate of the effect.

On the other hand, when comparing across municipalities, the picture is somewhat different. In regards to H1, when disaggregating the effect by time period, I do not find significance

of an effect, yet when aggregating the ATT coefficients from the intervention period I obtain weak significance of a positive effect of funding on the number of articles about CINs (see Table 4.2). To test the robustness of my results, I run the same model including the remainder articles about sport, obituaries and other, and find that the positive effect is still visible. The aggregated ATT coefficient of 4.92 indicates the average increase in the number of CINs articles for a treated municipality in the post intervention period, while the coefficient of 6.55 indicates the average increase in the number of articles overall for the treated municipalities. When looking at the individual time periods, as visible in Table 7.7 in Appendix A, no significant results are found in any of the three models, indicating that the significance of the first and second model only emerges when aggregating the ATTs of the different time periods in a joint estimate of the effect. The results indicate that the funding seem to have had a weak yet significant and positive impact on the journalistic production of the treated municipalities, and that reassuringly this effect is present even when only considering the most critically informative type of articles. In regards to H2, however, I find no support that the funding has positively impacted the number of articles about civic information and politics, suggesting that journalistic production did not particularly increase in those two categories.

5 Discussion

5.1 Classification

In relation to the first research question of this thesis, *RQ1: are local newspapers in blank spot municipalities providing articles that fulfill the critical information needs of their citizenry?*, the classification provides some guidance. We observe that the largest share of articles published by newspapers do indeed fulfill a critical information need, a statement that local newspapers act as source of vital information for a community. There are not substantial differences in label shares across blank spots and, where noticeable, attributable to inherent characteristics of the municipality. What the method and data do not capture, however, is how newspapers compare to other forms of media, such as radio, tv or hyperlocals, when it comes to CINs and local news. While being out of the scope of this thesis, this would provide an interesting expansion of this study. Another interesting extension of this research would be to classify articles belonging to both blank spots and non-blank spots municipalities, which were left out of this study given the focus on funding for blank spots. This would provide insights as to whether blank spots and non-blank spots are served differently by means of CINs, which could then shed light on whether the location of editorial offices plays a role in informing citizens on important local matters. On another note, it remains untested the reason for the noticed reduction in the number of articles, particularly visible for the CIN aggregated categories, overtime. Whether time trends and external events are the sole cause of why the number of articles decreases overtime, or whether there are additional underlying reasons, is an interesting emerging question.

The classification was done using Facebook's FastText model. While its overall architecture is described in the method section, what occurs within the model stages is somewhat of a black box, meaning that the motives behind the predictions are masked through the hidden layers of the algorithm's procedural steps. Nonetheless, the model's satisfactory accuracy and performance was considered a sufficiently positive sign that the model does what it is asked to do, that is, capturing the category of an article based on its semantic features. Yet, it is acknowledged that there is margin of error with the prediction accuracy. The usage of FastText supervised model to train and predict CINs is new to the literature, where the most similar approach is that of Weber, Andringa, and Napoli (2019), who used a dictionary of keywords based on the manual annotation research of Napoli, Weber, MCcollough, and Wang (2018) and applied it to articles' titles. The researchers classification accuracy method returned a score of 83%, which is higher than the present thesis score of 67%. Differences in the size of data available, as well as in the cultural contexts, make the comparison across methods challenging to make. However, one noticeable difference is that the category Other here captures around 10% of the full dataset (see Figure 4.3), whereas in Weber, Andringa, and Napoli (2019) it accounted for 52% of their sample. This could either indicate differences in coding choices, or differences in the type of journalism produced across the US and Swedish contexts. Finally, while this thesis adopts FastText, different classification approaches could be

tested, some that for instance use pre-trained word embeddings on the Swedish languages, where appropriate and available, or different predictors altogether.

For a classification task to have some value, its annotation needs to reliable, meaning that each label needs to be associated with the right answer (Fort, 2016). Due to the task of annotating often being a matter of interpreting text, it is common practice that several coders are asked to do the annotation. Then, metrics like Cohen's Kappa measure the differences in coding choices across coders and are used to evaluate the quality of the annotation. The annotation in this thesis was the work of a single individual, and I went through the annotation process twice, to solidify my own choices. Despite feeling overall consistent in my labelling, it is acknowledged that there could be a subjective bias as a result of being the sole coder of the hand-coded set. Furthermore, the choice of coding one thousand articles was done by compromising between the time needed to hand code and the size of the corpus. Most likely, a larger coded set would have yielded better classification results, by providing more examples to the model and by testing on a larger number of articles.

5.2 Statistical Analysis

The first set of models looked at the impact of funding on different dimension of journalistic production, in what I call a publisher-level analysis. The aim is to assess whether the funding has impacted the journalistic production of the funding receiving publishers in my sample, using as a counterfactual the untreated publishers within the same municipalities. Generally, insignificance is found in all three models, regardless of the dependent variable. This indicates that the funding has not impacted the volume of CINs article, nor of the entire journalistic production, and not even of the two most salient categories of articles traditionally relating to local newspapers: civic information and politics. The direction of the coefficient varies for the three models, with the first two having a positive coefficient, and the latter a negative, although all three coefficients are insignificant. Overall, while insignificant, the coefficients for the first two models seem to be trending upward, which sounds promising in relation to future estimations of the effect.

The second set of models looked instead at the impact of funding at the municipality-level, comparing all publishers in Bräcke and Vingåker, the two treated municipalities, to all publishers in Hallsberg and Nordanstig. What we see here is a slightly different picture. In regards to H2, I find no evidence of an effect of funding on the volume of articles about civic information and politics. However, when looking at the number of CINs articles and at the overall journalistic production, I find a small, yet significant, positive effect, suggesting that the funding has increased the volume of journalism in both those particularly relevant categories of information, and overall. Thus H1 is supported, and it highlights, jointly with H2, that the increase in production did not happen particularly in the categories of civic information and politics, which coincide with the type of journalism traditionally associated with the role of the local press. Why do we see an effect at the municipality-level but not at the level of the publisher? While untested in this thesis, the results raise some interesting opportunities for investigation. For example, it could be that the funding might have had a crowding out effect, by means of altering the journalism not just of the treated publisher, but also of the remainder publishers within the same municipalities.

All in all, the results lead us to believe that the funding has not had a significant effect on the volume of CINs, nor at the level of civic information and politics, for the treated publishers. When considering the entire journalistic production of a municipality where some publishers have received funding, however, we do see a small, yet significant, positive effect of funding

on the number of CINs articles. The results should be cautiously interpreted given the small sample sizes, with only a handful of funding-receiving publishers being present in the analyses. Further research could thus try to incorporate a larger number of municipalities and publishers in an analogous study, and see if the same direction and significance of the coefficients is still found. Thinking in terms of what we know of the funding so far, Tenor (2020) reported that for local newspapers over 20% of funding receiving municipalities in 2019 believed they saw an increase in both scope and quality of local journalism. This figure was against 16% who saw a negative change, and around 58% who saw no changes. The results found in this study somewhat complement this picture of mostly invisible and marginally positive changes. Due to only selecting two funded blank spots for the analysis, it would be interesting to obtain the individual responses of those municipalities from Tenor's (2020) survey, to see if what is found here matches the survey responses. Particularly, it could have been expected to see some changes in Bräcke, where Jämtlands Tidnining opened an editorial office in spring of 2020. However, given the Covid-19 pandemic, potential time lags in usage of funds, could all be playing a role in why no changes are observed.

Interesting considerations arise when thinking of alternative ways to setup this analysis. For example, an alternative way to run the publisher-level analysis is to compare the treated publishers to the top untreated publishers in the untreated municipalities. This might offer an alternative to the current analysis that accounts for differences across funding and non-funding receiving newsrooms within the same borders, as Table 3.3 showed that within the treated municipalities, the treated publishers produced on average a much larger amount of articles. Such a strategy would also account for a potential compensation effect or a crowding out effect, where the untreated publishers in the same municipality might publish less as a result of not being funded, or more (which is more plausible given the differential results between the publisher and municipality-level analyses). The current setup however, accounts for differences in external events that might have impacted differently the municipalities, and more generally unobserved differences across the selected blank spots.

5.2.1 Additional considerations

In the current statistical analysis the estimation of the effect of funding is purely based on whether funding was allocated or not, ignoring information on the amount received or how the funding was used. As shown in Figure 3.2, Bräcke and Vingåker received funding both in 2019 and 2020. Cumulatively, Vingåker ranked third by the amount of money received, while Bräcke ranked 9th. This makes them two highly funded municipalities, which might challenge the generalisability of the seen results across other blank spots. While Wallentin (2020) states that Bräcke's Jämtlands Tidning used the funding to open up a local editorial office (making Bräcke no longer a blank spot), it is not known specifically how the other publishers have used the funding, and thus such measures were not considered in this study.

The sampled municipalities were paired using a method that aimed at neutralizing the confounding effect of demographic characteristics associated with news production and consumption. However, news production is also determined by differences in newsrooms' abilities and strategies, dependent on manpower, financial resources and aim. Variables that would control for newsrooms structural differences were entirely left out of both the sampling strategy and the statistical analysis, due to challenges related to retrieving this type of information (hindered both by the lack of a reliable source and also by not knowing which newsrooms operate for a specific municipality). Regarding the sampling, their inclusion could have led to different pairs being formed, ones that minimise differences across the demographic characteristics of municipalities affecting news consumption, and the structural

aspects of the newsrooms operating for those. Adding structural characteristics of newsrooms would be an interesting extension of the statistical analysis, although it requires some considerations, for example the fact that the allocation of funding could have directly affected those, through the ability to hire a new journalist, or opening a new office, thus considering the role of these structural changes as mediators between the funding and the local journalism.

Another consideration lies in the unobserved factors that led the non-funded blank spots to be left unfunded. Given the multitude of non-blank spot municipalities which have received funding as well (see Appendix A for a list of those), a question remains unanswered as to why these areas were not allocated any funding. Potential explanations could be that no applications were submitted by the publishers covering them because of their greater financial capabilities, or because of lack of motivation to seek funding, or due to not meeting the criteria to be eligible for funding. Another motive would be that the region is not deemed sufficiently journalistically fragile, but given the focus of MPRT's funding creation on blank spots, this is deemed unlikely (MPRT, 2019). Regardless of the underlying reasons why those blank spots received no funding, it is acknowledged that there might be unobserved differences between regions or publishers in the control and treatment groups that relate to funding, which would pose as an endogeneity threat.

At the conceptual level, a blank spot is defined as a municipality lacking an editorial office within it. The definition thus implies that a municipality will be considered a blank spot even if a newsroom dedicated to covering it might be located just outside its borders. Additionally, a municipality could here be considered a blank spot even if its overall coverage across the country might be relatively high, by means of being covered by a major or large newsroom or several news outlets across the country. As mentioned previously, in Sweden there is no assembled data on which municipalities each news outlet covers specifically, making it challenging to address those issues. This thesis uses the best data at hand, the municipality in which editorial offices are located, to estimate blank spots. The geographical location of newsrooms is not used to assume that the municipality they reside in is what they cover - this inference is generally irrelevant for this study, as a proxy of which newspaper covers a municipality is obtained through the gathered publishers from the keyword search function on Retriever. On the other hand, obtaining the geographical location of news outlets could be combined with the extracted publisher names from a keyword search of a municipality name, to generate a mapping of which municipalities different news outlets talk about or aim at covering, creating a baseline for the identification of local news publishers that goes beyond the location of an editorial office. Particularly, this would help overcoming the challenges around studying blank spots, since by definition they do not have an editorial office within their borders. As media ownership tightens and centralized news production increases, linking titles with their area of coverage might become necessary in order to maintain an understanding of news coverage distribution in Sweden.

Regarding the quality of the data in capturing the local journalism of the selected municipalities, some limitations were presented, for example that some of the population of interest is missing, due to limitations imposed by the keyword search function. Future studies could attempt to implement further keywords, as done by Nygren, Leckner, and Tenor (2018), which were here not included given time limitations. Regarding noise in the data, the cutoff threshold used to remove irrelevant publishers was the best solution identified in the absence of documentation identifying key publishers for a given geographical region. Future research could also add a coding category indicating whether the article is about the local community, as done by Weber, Andringa, and Napoli (2019). A second classifier could then be trained to solely target whether an article is discussing something inherent to the local community or not, which would aim at capturing the salience of the municipality within the news article.

6 Conclusion

This research enriches the understanding of the state of local journalism in blank spots by introducing the Critical Information Needs framework to Swedish media literature. This was done through the automated classification of over fifty thousand articles written for six blank spot municipalities, by using Facebook's FastText supervised machine learning algorithm for text classification. The results show that a large share of local journalism articles do indeed fulfill a CIN, a reminder that papers hold an important informative role in their communities. The classification of articles in the CINs categories represents a step forward in our understanding of blank spots in Sweden, as suggested by Nygren and Schjærff Engelbrecht (2018), and, despite differences in context and study goals, it shows that unlike in the US (Napoli, Weber, MCcollough, & Wang, 2018), Swedish journalistically fragile municipalities are not really news deserts, and important news are still being delivered to local communities, even in the absence of an editorial presence.

The classification output was then used to test two hypotheses: to examine whether the Swedish state newly introduced financial subsidy for the production of local journalism has had any effect on the amount of CINs articles written for the sampled blank spots and to test whether the funding has impacted the traditionally most salient categories for local newspapers: civic information and politics. In relation to both H1 and H2, the results from two sets of difference-in-differences regression models show no significant effect of the funding at the publisher level, but a small yet significant and positive effect of the funding at the municipality for the CINs articles, which is although not reflected when only looking at the categories of civic information and politics.

All in all, the results of this thesis show that the subsidy has not significantly altered the picture of journalistic content type for the funded publishers. At the municipality level, it has, whenever noticeable, only marginally positively impacted the local journalism of these blank spots, although reassuringly the effect can be seen in the categories of information that fulfill citizens CINs. The difference in results between the publisher and municipal-level of analysis sparks interest in why we see an effect when we take into account the unfunded publishers in the same municipality as the funded ones, suggesting perhaps a crowding out effect of the funding for the untreated outlets. At the publisher-level, the results generate questions about the potential alternative ways the funding might have helped individual publishers and reporters sustain and strengthen the local journalism of these blank spots, or which areas of coverage have been given more attention since the funding has been received. Particularly, it appears plausible that the funding has either been used in ways that did not directly aim at impact the journalistic production of civic information and political stories, or that it might take time before the effect will actually take place. Overall, this thesis offers a first quantitative assessment of the effect of the funding on local journalism, currently untested in the literature.

On a different note, this thesis aims at stimulating research on blank spots by showcasing the viability of using computational text analysis to analyse large amounts of texts, and in its completion it generates a multitude of exciting opportunities for extension. For example, the CINs framework could be expanded by coding whether a story was original and about the local community, as done by CIN researchers Napoli, Weber, MCcollough, and Wang (2018), helping the Swedish news media ecological system. This could particularly offer interesting insights on the role of local newspapers in relation to other mediums such as citizen journalists and local television, as similarly attemped by Mahone, Wang, Napoli, Weber, and McCollough (2019). Additionally, the CINs framework could be used to compare differences in the state of local journalism between blank spots and non-blank spots, offering a stance in the debate over the geographical location of editorial offices that takes into account the traditionally informative role of local newspapers and the democratic information needs of citizens. In relation to the funding, more research is needed to offer an understanding of the efficacy of state subsidies on the strengthening and sustenance of local journalism, one that takes into account the measures taken by the subsidised news outlets, beyond the opinions of municipality officials (Tenor, 2020). The results of this study complement the overall literature's understanding of local news outlets being providers of vital local information, although this is done irrespective of other mediums. Thus, the results should only be interpreted as a reassurance that blank spots do indeed receive critical information, despite not having an editorial presence situated in the municipality, and they do so in a similar fashion regardless of whether publishers within them have or not received the state funding for local journalism.

7 Appendix A

This thesis was accomplished through the usage of statistical software, namely the R programming language within the RStudio environment. Listed below are the specifications of the different components, as well as the packages used. The code used to carry out this thesis is located in the folder Appendix B, within which are present the R and RMarkdown scripts that constitute this project, and the files necessary for the analysis or generated throughout the analysis.

- Software: R
- R version 4.0.4 (2021-02-15) – "Lost Library Book"
- RStudio version 1.4.1106, created by: RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
- Platform: x86_64-w64-mingw32/x64 (64-bit)
- Running under: Windows 10 x64
- An additional RStudio image Server was used to carry out the SML section of this thesis, due to limitations in the computation power of the main machine used in this study.

Packages

- Stargazer, for table formatting: Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
- FastTextR, for the fastText classification: Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of tricks for efficient text classification.arXiv preprint arXiv:1607.01759
- CMD (not used in the paper, but found in the appendix). Stoltz, Dustin S., and Marshall A. Taylor. 2019. "Concept Mover's Distance." Journal of Computational Social Science 2(2):293-313.
- did, for the DID analysis. Callaway, Brantly, and Pedro H.C. Sant'Anna. Difference-in-differences with multiple time periods. Forthcoming at the Journal of Econometrics (2020).

7.1 Critical Information Needs

An elungated explanation of each critical information need can be found here. From Friedland, Napoli, Ognyanova, Weil, and Wilson III (2012):

1. Emergencies and Public Safety: Individuals, neighborhoods, and communities need access to emergency information on platforms that are universally accessible and in languages understood by the large majority of the local population, including information on dangerous weather; environmental and other biohazardous outbreaks; and public safety threats, including terrorism, amber alerts, and other threats to public order and safety. Further, all citizens need access to local (including neighborhood) information on policing and public safety.
2. Health: All members of local communities need access to information on local health and healthcare, including information on family and public health in accessible languages and platforms; information on the availability, quality, and cost of local health care for accessibility, lowering costs, and ensuring that markets function properly, including variations by neighborhood and city region; the availability of local public health information, programs, and services, including wellness care and local clinics and hospitals; timely information in accessible language on the spread of disease and vaccination; timely access to information about local health campaigns and interventions.
3. Education: Local communities need access to information on all aspects of the local educational system, particularly during a period when local education is a central matter for public debate, decision-making, and resource allocation, including: the quality and administration of local school systems at a community-wide level; the quality of schools within specific neighborhoods and geographic regions; information about educational opportunities, including school performance assessments, enrichment, tutoring, afterschool care and programs; information about school alternatives, including charters; information about adult education, including language courses, job training, and GED programs, as well as local opportunities for higher education.
4. Transportation Systems: All members need timely information about local transportation across multiple accessible platforms, including: information about essential transportation services including mass transit at the neighborhood, city, and regional levels; traffic and road conditions, including those related to weather and closings; timely access to public debate on transportation at all layers of the local community, including roads and mass transit.
5. Environment and Planning: Local communities need access to both short and long-term information on the local environment, as well as planning issues that may affect the quality of lives in neighborhoods, cities, and metropolitan regions, including; the quality of local and regional water and air, timely alerts of hazards, and longer term issues of sustainability; the distribution of actual and potential environmental hazards by neighborhood, city region, and metropolitan area, including toxic hazards and brownfields; natural resource development issues that affect the health and quality of life and economic development of local communities; information on access to environmental regions, including activity for restoration of watersheds and habitat, and opportunities for recreation.

6. Economic Development: Individuals, neighborhoods, and communities need access to a broad range of economic information, including: employment information and opportunities within the local region; job training and retraining, apprenticeship, and other sources of reskilling and advancement; information on small business opportunities, including startup assistance and capital resources; information on major economic development initiatives affecting all local levels.
7. Civic Information: Communities need information about major civic institutions, nonprofit organizations, and associations, including their services, accessibility, and opportunities for participation in: libraries and community-based information services; cultural and arts information; recreational opportunities; nonprofit groups and associations; community-based social services and programs; and religious institutions and programs.
8. Political Life: In a federal democracy, citizens need information on local, regional, and county candidates at all units of governance, including: information on elected and voluntary neighborhood councils; school boards; city council and alder elections; city regions; and county elections; timely information on public meetings and issues, including outcomes; information on where and how to register to vote, including requirements for identification and absentee ballots; information on state-level issues where they impact local policy formation and decisions.

7.2 Kommundatabasen criteria for newsroom to fit in the count

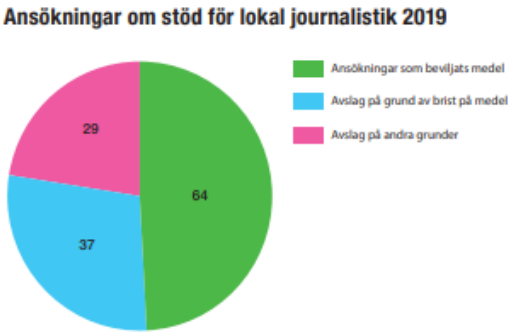
Here below is reported the criteria for newsrooms to feature the Kommundatabasen website. Their choices have an impact on what I will recognise as a blank spot or not, thus it is important to comprehend what is being included and what is being cut out.

> The database includes titles that are local, regular, and at least to some extent contain editorial material with a general coverage of the local community. Publications that are published or updated irregularly or less frequently than monthly are not included here, nor are publications that completely lack local coverage focused on a specific geographical area. Only titles that have their own platform (paper newspaper, site, etc) are included. This means that we exclude, for example, Facebook pages and newsletters. In total, there are almost 1,000 editorial offices in the database, both current and those that have been closed down.

7.3 Local Journalism Financial Basic Criteria and other information about funding

From the Årsredovisning Report 2019 (p.17):

> Myndigheten har under året berett tillfälle att ansöka om stöd för lokal journalistik för första gången. Sammanlagt har 130 ansökningar från 34 företag kommit in till myndigheten. Det totala sökta beloppet uppgick till 103,8 miljoner kronor.



Figur 8. Antal inkomna ansökningar om stöd för lokal journalistik i områden med svag bevakning och antal avslag år 2019. Källa: Myndighetens databas.

Figure 7.1: Image from the Årsredovisning Report 2019. Applications outcomes for local journalism funding. Data from MPRT

From the Årsredovisning Report 2020 (p.22):

> Myndigheten har under året berett tillfälle att ansöka om stöd för lokal journalistik för andra gången. På grund av covid-19-pandemin utökade regeringen anslaget för stöd till lokal journalistik. Regeringen höjde även taket för hur mycket medel som kan beviljas i stödet för lokal journalistik från en miljon per område till två miljoner per område. Sammanlagt har 231 ansökningar från 57 företag kommit in till myndigheten. Det totala sökta beloppet uppgick till 265,8 miljoner kronor. Totalt beviljades 192 ansökningar medel och 121,4 miljoner kronor fördelades. De sökta beloppen sattes ner utifrån en prioriteringsordning för att uppnå en så stor geografisk spridning som möjligt. Stödet fördelade sig över 181 kommuner i 20 län. På grund av brist på medel avslogs 15 ansökningar. Ytterligare 24 ansökningar avslogs eller skrevs av andra skäl, exempelvis att ansökningarna inte var kompletta, att sökanden återkallade ansökan eller att det aktuella området inte kunde ses som svagt bevakat. Under 2020 har 1,9 miljoner kronor i stöd för lokal journalistik återbetalats efter mediestödsnämndens beslut om återkrav.

An elungated explanation of the basic criteria for a newsroom to be considered for the reception of funds can be found here. From MPRT (2019):

> You can apply for support for initiatives for journalistic coverage, which means that a news medium starts, expands or maintains journalistic coverage of an area that lacks or has weak coverage. Support for local journalism is granted per year and can amount to a maximum of two million kronor per area. The support can be distributed between up to three actors in the same area, if it promotes the journalistic coverage of the area. In some cases, support can be granted over several years, a maximum of three years. Support for local journalism can be sought in collaboration between several different general news media. Those who are part of the collaboration can apply for support individually, jointly or appoint an applicant as responsible for the initiative.
>
> In order to receive support for local journalism, your newspaper needs to meet a number of criteria. These are: The medium must be a general news medium The

medium must meet the basic requirements for a news medium The media shall promote the accessibility of the editorial content for people with disabilities The media must submit a technical report or a media support certificate

To be seen as a general news medium, the medium's primary task must be to conduct news coverage. The medium must also have an editorial content consisting of regular and comprehensive news coverage that expresses a wide range of topics and perspectives as well as an examination of the basic events for democracy.

Support can be provided regardless of the content and distribution form of the medium. This means that the content can consist of e.g. text, image, sound or moving image and support can be given to e.g. newspapers, web-based media, radio or television. Support can be provided to both paid media and free media.

To be eligible for support, the general news media must: Have an editorial content that makes up at least 50 percent of the entire content. Have exclusive, self-produced editorial content consisting of news coverage and review of at least 20 percent of the entire content. Be publicly available. Have your own title with independent editorial resources. Have a responsible publisher. Follow good media ethical practice. Promote accessibility for people with disabilities. Be aimed at a Swedish target group. Published or broadcast at least 45 times a year (high regularity). Have at least 1,500 regular users (good user base). It is circumstances during the year preceding the application that form the basis for the assessment.

There are reliefs, for example for general news media aimed at the national minorities and for newly started general news media.

Requirements - promotion of accessible content A prerequisite for being granted support is that the medium actively works to improve accessibility to the editorial content for people with disabilities. A starting point should be that universal design is sought as far as possible, ie. design so that it can be used to the greatest possible extent by everyone without the need for adaptation or special design. The work must be based on current standards for accessibility in digital publishing.

The size of the general news medium and the economic and technical conditions must be taken into account when assessing whether this requirement is met. You can read more about how you can make your content available under our pages on Accessibility.

Requirements - technical report or media support certificate General news media seeking support must submit information on good user base, high regularity, scope of editorial content, share of editorial content and share of exclusive, self-produced editorial content.

To measure whether the requirements are met, a report must be submitted. A special certificate can be produced by Media Audit at Kantar Sifo.

An alternative is for the news media to produce and submit a technical report themselves. The technical report consists of traffic reports, editions and a so-called edition protocol. Instructions for reporting can be found in the Instructions for technical accounting prior to applying for media support. Supplementary information on traffic measurements can be found under Approved third-party tools.

A technical report must be certified by a qualified auditor. Instructions and template for audit can be found in Template for audit form.

7.4 Overview of funding reception for all Swedish municipalities

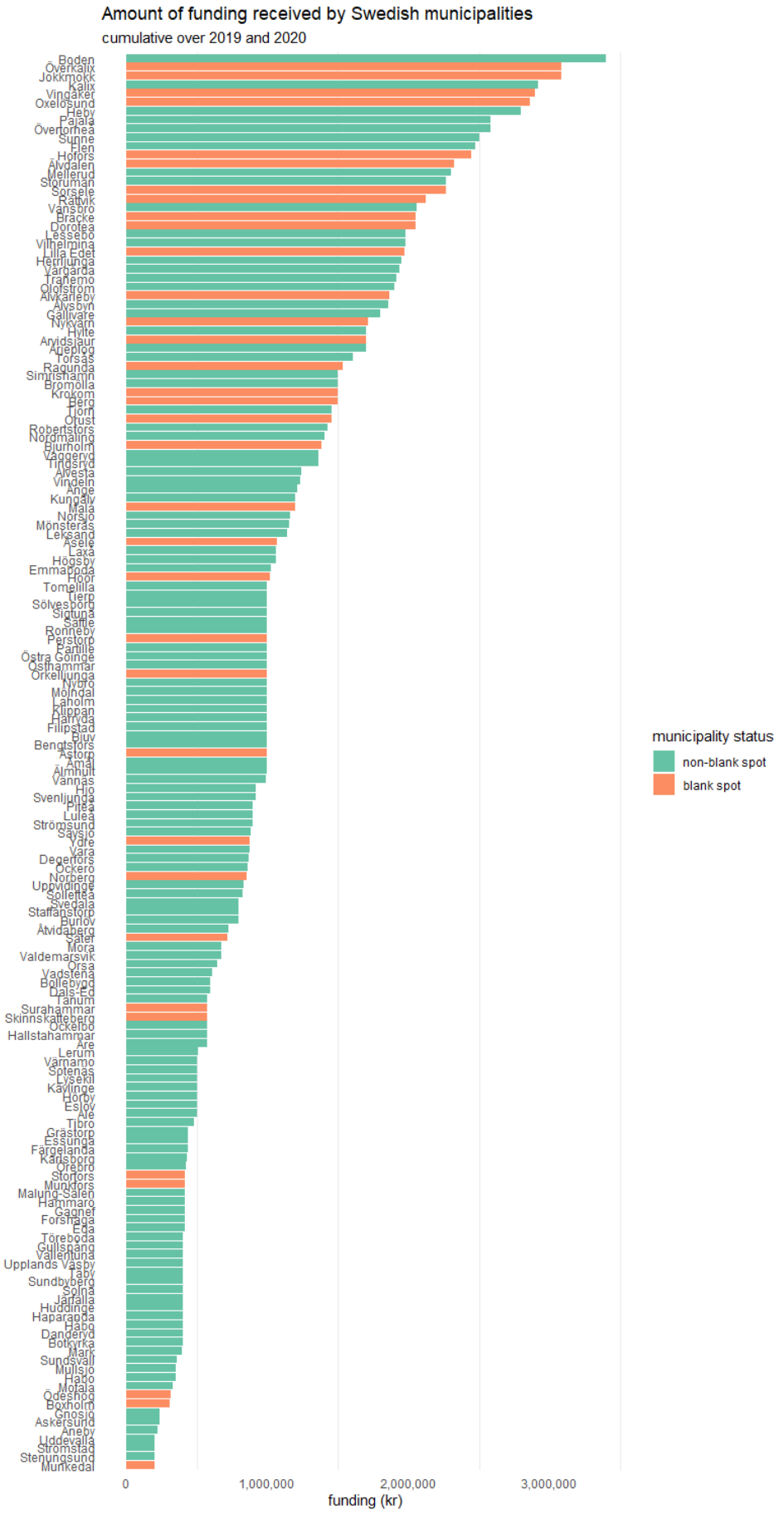


Figure 7.2: Amount of funding received by all funding receiving municipalities. Data from Myndigheten för Press, Radio, och TV

7.5 Summary Statistics of Demographic Variables

Table 7.1: Summary Statistics of Demographic Variables across all 290 municipalities

Statistic	Mean	St. Dev.	Min	Max
Population (n)	35,276.5	73,307.9	2,450	962,154
People with post secondary education (%)	33.3	9.7	19.9	73.2
Women (%)	49.2	0.8	46.6	51.1
Average wage (mkr)	6,163.0	15,330.4	308.4	218,132.8
Average age (years)	43.3	2.6	36.4	49.8
Foreign-born (%)	15.4	6.3	6.5	42.1
Last general election turnout (%)	87.2	2.5	72.8	93.9

7.6 RKA'S KOLADA similarity score calculation

This section describes the two-step process to obtain the similarity score on the Kolada website. First, the variables are normalised, to neutralise their difference in magnitude. Normalisation of each variable is obtained through the following formula:

$$y_j^i = \frac{v_j^i}{|per_j^{65} - per_j^{35}|}$$

Where
y = The magnitude of the variable j
v = The magnitude of the variable j
per_j^{65} = percentile x for variable j

Following the normalisation of the variables, an index is calculated for each municipality i in the country. The index gives a score of dissimilarity, thus a score of 0.2 indicates higher similarity than a score of 0.5. The index for a given municipality is generated through the formula:

$$index_i = \sum_{j=1}^{m} \mathbf{w}_j |y_j^{fokus} - y_j^i| \tag{7.1}$$

Following this process, the most similar funded blank spot to each of the four selected non-funded blank spots was selected. Some considerations are made in relation to the sampling. The normalization of the variables used in the Kolada matching procedure, followed by the creation of an index where those variables were given equal weight, could have given too much emphasis on variables where the differences between municipalities would have been minimal (e.g. percentage of women), while reducing the importance of differences across other variables (e.g. average wage). The consequences are that the pairing might be overly skewed towards differences that are relatively less important and substantial than others. It would be an interesting application for future researchers attempting a similar sampling strategy, to test different variables, or assign different weights to them.

7.7 Funding Allocation within the selected sample

The total number of funding receiving publishers in the selected sample is five. In Älvdalen, the newspaper Dalabygden was awarded 853875 kr in 2019 and 800 000 kr in 2020 to cover Älvdalen alongside three additional municipalities. In Bräcke, Jämtlands Tidining was given 400000 kr in 2019 and 440000 kr in 2020, while Östersunds-Posten received 576000 kr in 2020 and Sundsvalls Tidning 637874 kr (to use for both Bräcke and Ånge) in 2019. Finally, the municipality of Vingåker received 400000 kr through Sörmlandsbygden in 2019 (also used for the municipality of Oxelösund), which increased to 500000 kr in 2020, and 800000 kr in 2019 through Katrineholms-Kuriren, which increased to 1198354 kr in 2020.

7.8 An alternative to FastText: Concept Mover's Distance

An important step in a SML task is the choice of predictors. What, in a given article, makes it so that it fulfills a particular critical information need? The annotation process provided some guidance in this matter. Particularly, it became apparent that for each CIN labelled article there was a set of identifiable words that signalled the association of the article to its label. To capture those label-signalling words, I used the Term Frequency Inverse Document Frequency (TF-IDF), which calculates the importance of words in a document by measuring the inverse proportion of the frequency of a document's word to the share of documents it appears in (Ramos et al., 2003). The TF-IDF is calculated following the formula:

$$tfidf_{i,d} = tf_{i,d} \cdot idf_i \tag{7.2}$$

Where
$tf_{i,d}$ = the term frequency of a word in the documents of a given label, giving an estimate of how frequent a word is within a document or group of documents
$idf_{i,d}$ = the logarithm of the number resulting by dividing the number of documents by the number of documents containing a given word, giving an estimate of how rare or common the word is across all documents in the corpus.

The top 20 TF-IDF words were selected to form a dictionary of keywords describing each of the eleven categories. The 220 keywords are then selected as predictors of the category of an article. Because across documents words might change slightly, the particular TF-IDF extracted from the annotated set might not be representative of what they would look like if extracted from the full corpus, which was not possible because of it being unlabelled. To overcome the penalty associated with the fact that one article might use similar words about a concept, such as politics, but not precisely the same ones as most other articles about politics in the corpus, I here used a novel metric proposed by Stoltz and Taylor (2019), known as Concept Mover's Distance (CMD). CMD is a vector of values that describe the degree of similarity between documents and some focal concepts. CMD builds upon Relaxed Word Mover's Distance (RWMD), a method that uses the word embeddings of the words within a document to estimate its distance from other documents in a corpus. Word-embeddings are mathematical representations of a word position in a hyper-dimensional space, typically on 300 dimensions. Words that appear close to each other on most positions within the word embedding vector space, tend to be related in semantic context and meaning. Word embeddings thus prove to be very helpful in tasks that want to capture meaning or concepts, and allow for more flexibility in the lexicon used to discussed those. The word embedding vector

of a particular word can be calculated from one own's data or it can be downloaded in the form of a pre-trained model. Rather than estimating the distance between documents, CMD has the peculiar feature of measuring the distance, or "similarity", between a document and a keyword. This is done by means of calculating the cosine distance between each word in a document and the keyword of interest. The distance, measured by means of word embeddings location, is then multiplied by the relative frequency of the word in a document. The resulting list of distances is averaged across the words in the document, returning a final distance list that equals the CMD for that document and the focal keyword.

CMD inspired this thesis by offering a way of measuring the distance between the article's documents and a particular critical information need, by means of allowing for flexibility in lexical choices within different articles, and using keywords strongly associated with each label as the base for the prediction. Given CMD requires word embeddings, I availed of Facebook's FastText pre-trained model for the Swedish language, which calculates the hyper-dimensional position of words from a large corpus of text extracted from Wikipedia and Common Crawl (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). While there might be differences in the usage of language between Swedish news articles and the data FastText uses to construct the word embeddings, those are not expected to be susbtantial or to generate very different word embeddings. The word embeddings vectors for each word in a vocabulary consisting of the TF-IDF keywords and the words in the annotated articles were then used to obtain a metric measuring the distance between each article and each of the keywords.

For each document in the annotated sample, excluding stopwords, the CMD to each of the 220 keywords was calculated. Interestingly, some words that appeared in the top 20 TF-IDF did not have a word embedding vector in the model, and were dropped. Those included names of geographical places and words associated with the pandemic, which started after the release of the word embedding models in 2018 (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). A random forest classification model, a support vector machine, and a multinomial logistic regression model were thereafter trained on 80% of the articles in the annotated set. No pre-processing steps were required for the predictors, that is the CMD values for each keyword to each labelled document. The random forest model performed best, and it was thereafter tuned, following a grid search, for the best hyperparameters. The best performing model on the basis of the ROC AUC value was selected. The model was trained using 10-fold cross validation, a procedure that further divides the train data in 10 combinations of train and test batches, to validate the model. Finally, once the model was trained, it was used on the CMD matrix of distances for all documents in the full dataset and the keywords, and that way classified the full dataset.

The accuracy of predictions for the test set was around 60%, slightly lower than the FastText model used in this paper. While this thesis used FastText because of better results in the evaluation metrics, CMD seemed to offer a valid alternative to the model chosen in this paper.

7.9 FastText Model Specifications

Here below are presented the specification of the fastText classifier.

- loss function: softmax
- learning rate: 0.5
- learn update: 100

- word vector size: 50
- window size: 5
- epoch: 7000
- minimum count: 1
- minimum label count: 0
- negative: 5
- maximum length of n-grams: 10
- buckets: 10000000
- minimal n-gram length: 3
- maximal n-gram length: 6
- threads: 4
- threshold: 10000
- label: "__label__"
- verbose: 0
- pretrained vectors: ""
- output: ""
- save output: FALSE
- seed: 0
- qnorm: FALSE
- qout: FALSE
- cutoff: 0
- dsub: 2
- autotune validation file: ""
- autotune metric: "f1"
- autotune prediction: 1
- autotune duration: 300
- autotune model size: ""
- attr(,"class"): "ft_control" "list"

7.10 Panel Data Summary Structure

Here below is presented an overview of the subjects within the panel data structures used for the statistical analyses.

Table 7.2: Overview of subjects in the publisher-level analysis

	publisher	municipality	Time Points
1	Aftonbladet	Bräcke	12
2	Aftonbladet	Vingåker	12
3	Alingsås Tidning	Bräcke	12
4	Dagens Nyheter	Bräcke	12
5	Dagens Nyheter	Vingåker	12
6	Dagens Samhälle	Bräcke	12
7	Dala-Demokraten	Vingåker	12
8	Enköpings-Posten	Vingåker	12
9	Eskilstuna-Kuriren	Vingåker	12
10	Eskilstuna Kuriren	Vingåker	12
11	Expressen	Bräcke	12
12	Expressen	Vingåker	12
13	Extra KK	Vingåker	12
14	Folkbladet	Vingåker	12
15	Göteborgs-Posten	Bräcke	12
16	Hallands Nyheter	Vingåker	12
17	Helsingborgs Dagblad	Bräcke	12
18	Helsingborgs Dagblad	Vingåker	12
19	Jämtlands Tidning	Bräcke	12
20	Katrineholms-Kuriren	Vingåker	12
21	Länstidningen Östersund	Bräcke	12
22	Nerikes Allehanda	Vingåker	12
23	ÖP magasin	Bräcke	12
24	Östgöta Correspondenten	Vingåker	12
25	Provinstidningen Dalsland	Bräcke	12
26	Skaraborgs Läns Tidning	Bräcke	12
27	Södermanlands Nyheter	Vingåker	12
28	Sörmlandsbygden	Vingåker	12
29	SR P4 Sörmland	Vingåker	12
30	Sundsvalls Tidning	Bräcke	12
31	Svenska Dagbladet	Bräcke	12
32	Sveriges Radio Jämtland	Bräcke	12
33	Sveriges Radio Sörmland	Vingåker	12
34	SVT Nyheter	Bräcke	12
35	SVT Nyheter	Vingåker	12
36	SVT Nyheter Jämtland	Bräcke	12
37	Tidningen Härjedalen	Bräcke	12
38	TT Nyhetsbyrån	Bräcke	12
39	TT Nyhetsbyrån	Vingåker	12

Table 7.3: Overview of subjects in the municipality-level analysis. Göteborg-Posten the one missing in the Politics and Civic Information Model - 1 of 2

	municipality	publisher	n
1	Bräcke	Aftonbladet	12
2	Bräcke	Alingsås Tidning	12
3	Bräcke	Dagens Nyheter	12
4	Bräcke	Dagens Samhälle	12
5	Bräcke	Expressen	12
6	Bräcke	Göteborgs-Posten	12
7	Bräcke	Helsingborgs Dagblad	12
8	Bräcke	Jämtlands Tidning	12
9	Bräcke	Länstidningen Östersund	12
10	Bräcke	ÖP magasin	12
11	Bräcke	Provinstidningen Dalsland	12
12	Bräcke	Skaraborgs Läns Tidning	12
13	Bräcke	Sundsvalls Tidning	12
14	Bräcke	Svenska Dagbladet	12
15	Bräcke	Sveriges Radio Jämtland	12
16	Bräcke	SVT Nyheter	12
17	Bräcke	SVT Nyheter Jämtland	12
18	Bräcke	Tidningen Härjedalen	12
19	Bräcke	TT Nyhetsbyrån	12
20	Hallsberg	Aftonbladet	12
21	Hallsberg	Bblat	12
22	Hallsberg	Dagens Nyheter	12
23	Hallsberg	Dala-Demokraten	12
24	Hallsberg	Dalarnas Tidningar	12
25	Hallsberg	Eskilstuna-Kuriren	12
26	Hallsberg	Expressen	12
27	Hallsberg	Göteborgs-Posten	12
28	Hallsberg	GT	12
29	Hallsberg	Katrineholms-Kuriren	12
30	Hallsberg	KT-Kuriren	12
31	Hallsberg	Länsposten	12
32	Hallsberg	Mariestads-Tidningen	12
33	Hallsberg	Motala, Vadstena Tidning	12
34	Hallsberg	Nerikes Allehanda	12
35	Hallsberg	Nya Wermlands-Tidningen	12

Table 7.4: Overview of subjects in the municipality-level analysis. Göteborg-Posten the one missing in the Politics and Civic Information Model - 2 of 2

	municipality	publisher	n
36	Hallsberg	Provinstidningen Dalsland	12
37	Hallsberg	Södermanlands Nyheter	12
38	Hallsberg	Svenska Dagbladet	12
39	Hallsberg	Sveriges Radio Örebro	12
40	Hallsberg	SVT Nyheter	12
41	Hallsberg	SVT Nyheter Örebro	12
42	Hallsberg	TT Nyhetsbyrån	12
43	Hallsberg	Värmlands Folkblad	12
44	Hallsberg	Vestmanlands Läns Tidning	12
45	Nordanstig	Arbetarbladet	12
46	Nordanstig	Gefle Dagblad	12
47	Nordanstig	Hela Hälsingland	12
48	Nordanstig	Hudiksvalls Tidning	12
49	Nordanstig	Ljusdals-Posten	12
50	Nordanstig	Ljusnan	12
51	Nordanstig	Söderhamns-Kuriren	12
52	Nordanstig	Sundsvalls Tidning	12
53	Nordanstig	Sveriges Radio Gävleborg	12
54	Nordanstig	SVT Nyheter Gävleborg	12
55	Vingåker	Aftonbladet	12
56	Vingåker	Dagens Nyheter	12
57	Vingåker	Dala-Demokraten	12
58	Vingåker	Enköpings-Posten	12
59	Vingåker	Eskilstuna-Kuriren	12
60	Vingåker	Eskilstuna Kuriren	12
61	Vingåker	Expressen	12
62	Vingåker	Extra KK	12
63	Vingåker	Folkbladet	12
64	Vingåker	Hallands Nyheter	12
65	Vingåker	Helsingborgs Dagblad	12
66	Vingåker	Katrineholms-Kuriren	12
67	Vingåker	Nerikes Allehanda	12
68	Vingåker	Östgöta Correspondenten	12
69	Vingåker	Södermanlands Nyheter	12
70	Vingåker	Sörmlandsbygden	12
71	Vingåker	SR P4 Sörmland	12
72	Vingåker	Sveriges Radio Sörmland	12
73	Vingåker	SVT Nyheter	12
74	Vingåker	TT Nyhetsbyrån	12

7.11 Descriptive Statistics of text classification output, aggregated by CIN categories, and non-CIN categories

Table 7.5: Summary Statistics of CIN and non-CIN articles overtime, aggregated by yearly quarter, across publishers (it shows the CIN articles account on average for 75% of the total journalistic production

Category	Min	1st Qu.	Median	Mean	3rd Qu.	Max
non-CIN	29.50	45.88	54.62	53.85	62.50	73.00
CIN	89.25	154.81	169.25	173.74	193.38	268.75

7.12 Statistical Analysis Results

7.12.1 Publisher-Level Analysis

Table 7.6 the ATT coefficients for each time period, including the pre-intervention ones, which can be used as a pre-test of the parallel trend assumption.

7.12.2 Municipality-Level Analysis

Table 7.7 the ATT coefficients for each time period, including the pre-intervention ones, which can be used as a pre-test of the parallel trend assumption.

7.12.3 Results of models where I remove Göteborgs-Posten

Because in model one and two I have one more publisher than in the third model, when analysing at the municipality level, I here test the same two models without that publishers, to see if my results are consistent. The results show the same direction and significance as in my models presented in the main paper, suggesting comparability across models despite a slightly different sample size between model one and two, and model three.

Table 7.6: Publisher-level analysis: ATT coefficients for the individual time points. Robust standard errors are reported

CINs				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	-34.0857	38.0747	-107.4686	39.2971
2018.3	8.7214	19.0946	-28.0802	45.5231
2018.4	-5.0786	6.7101	-18.0111	7.8540
2019.1	8.7929	5.5209	-1.8477	19.4335
2019.2	-46.5357	50.7500	-144.3481	51.2767
2019.3	-11.8500	15.3984	-41.5279	17.8279
2019.4	14.6571	15.5328	-15.2799	44.5941
2020.1	24.8857	14.7885	-3.6167	53.3881
2020.2	18.6571	11.7564	-4.0015	41.3158
2020.3	20.4929	13.4758	-5.4795	46.4653
2020.4	23.5571	14.5039	-4.3968	51.5111
All articles				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	-45.4500	64.3555	-160.9838	70.0838
2018.3	17.4286	26.2935	-29.7746	64.6318
2018.4	-16.4571	18.5997	-49.8481	16.9338
2019.1	5.8286	4.7661	-2.7278	14.3849
2019.2	-52.6286	76.4342	-189.8465	84.5894
2019.3	-19.5143	25.2490	-64.8424	25.8138
2019.4	20.8786	21.9906	-18.5999	60.3571
2020.1	26.3357	16.4427	-3.1829	55.8543
2020.2	15.2571	8.9803	-0.8647	31.3790
2020.3	21.4929	16.4455	-8.0307	51.0165
2020.4	38.1643	35.9924	-26.4509	102.7794
Civic Info and Politics				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	-14.8857	13.7834	-40.8569	11.0854
018.3	18.9571	28.0634	-33.9210	71.8353
2018.4	-18.2643	23.4137	-62.3813	25.8527
2019.1	-3.8357	6.5707	-16.2165	8.5451
2019.2	-26.7643	29.2017	-81.7874	28.2588
2019.3	-3.2857	3.0530	-9.0382	2.4668
2019.4	6.9214	9.7508	-11.4513	25.2942
2020.1	2.8714	4.6077	-5.8106	11.5535
2020.2	-3.9571	12.5401	-27.5857	19.6714
2020.3	-4.5714	5.8669	-15.6260	6.4832
2020.4	-6.4786	10.7243	-26.6857	13.7285

Table 7.7: Municipaliy-level: ATT coefficients for individual time periods. Robust standard errors are reported

CINs				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	1.1326	5.7739	-12.3833	14.6485
2018.3	8.1231	5.3334	-4.3617	20.6079
2018.4	-13.5780	9.7368	-36.3705	9.2144
2019.1	5.5018	4.5837	-5.2280	16.2317
2019.2	-2.8015	6.7802	-18.6729	13.0700
2019.3	8.4703	7.6110	-9.3459	26.2865
2019.4	0.2176	2.3346	-5.2475	5.6827
2020.1	4.7524	2.7060	-1.5819	11.0866
2020.2	7.4278	3.4204	-0.5789	15.4346
2020.3	5.2447	2.5716	-0.7752	11.2645
2020.4	3.4161	2.8086	-3.1585	9.9907
All articles				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	2.9385	7.2538	-13.3375	19.2144
2018.3	9.3018	5.5291	-3.1044	21.7080
2018.4	-20.0593	12.4355	-47.9622	7.8435
2019.1	6.6549	4.6182	-3.7075	17.0174
2019.2	-3.2549	7.9713	-21.1408	14.6309
2019.3	9.1414	9.1216	-11.3256	29.6083
2019.4	0.9414	3.3502	-6.5758	8.4586
2020.1	6.2440	3.0991	-0.7098	13.1977
2020.2	9.6220	5.4241	-2.5485	21.7925
2020.3	3.9187	3.0458	-2.9154	10.7528
2020.4	9.4286	5.2398	-2.3284	21.1856
Civic Info and Politics				
Time	ATT(g,t)	Std. Error	[95% Conf. Int.]	
2018.2	-1.0234	2.3591	-6.3632	4.3164
2018.3	5.5754	3.7022	-2.8043	13.9551
2018.4	-5.8296	3.7441	-14.3041	2.6449
2019.1	2.0158	2.4191	-3.4598	7.4915
2019.2	-4.6719	3.8511	-13.3888	4.0449
2019.3	3.2428	3.2340	-4.0772	10.5629
2019.4	1.6267	1.6539	-2.1168	5.3702
2020.1	2.9231	1.6807	-0.8811	6.7272
2020.2	3.4925	2.7197	-2.6634	9.6483
2020.3	3.7202	3.9365	-5.1898	12.6302
2020.4	1.3288	2.1016	-3.4282	6.0858

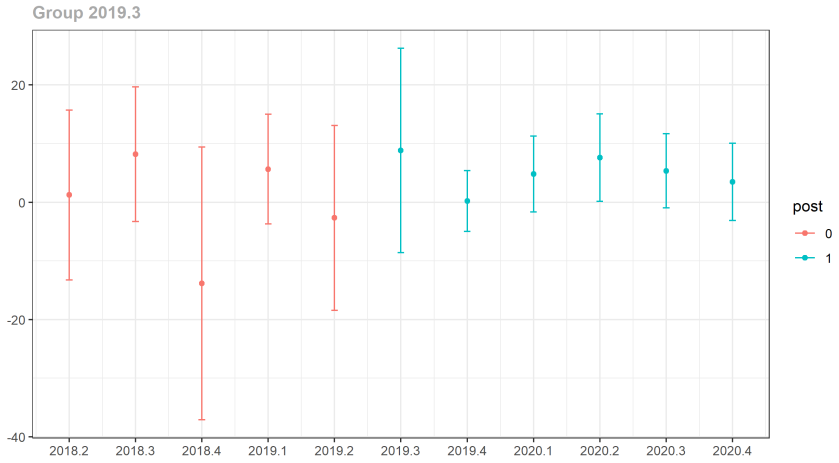


Figure 7.3: ATT coefficients showing the impact of funding at each year quarter on the number of CIN articles, at the municipality level, excluding Göterborgs-Posten from the analysis

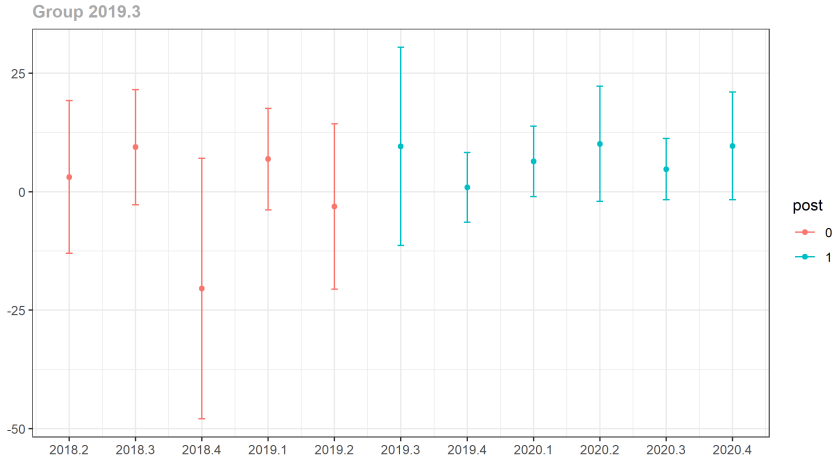


Figure 7.4: ATT coefficients showing the impact of funding at each year quarter on the number of overall articles, at the municipality level, excluding Göterborgs-Posten from the analysis

Table 7.8: Averaged ATT coefficients for the overall intervention time period for two models that exclude the publisher Göteborgs-Posten. Robust standard errors are reported

MUNICIPALITY-LEVEL			
CINs			
ATT	Std. Error	[95% Conf. Int.]	
5.04	2.38	0.36	9.73 *
All articles			
ATT	Std. Error	[95% Conf. Int.]	
6.88	2.87	1.24	12.52 *
Signif. codes: '*'			
Control Group: Never Treated			
Est. Method: Doubly Robust			

Bibliography

Abernathy, P. M. (2018). *The expanding news desert*. Center for Innovation; Sustainability in Local Media, School of Media; Journalism, University of North Carolina at Chapel Hill.

Andersson, U. (2013). From broadsheet to tabloid: Content changes in swedish newspapers in the light of a shrunken size. *Observatorio (OBS*)*, *7*(4).

Barbera, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42.

Burggraaff, C., & Trilling, D. (2020). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, *21*(1), 112–129.

Callaway, B., & Sant'Anna, P. H. C. (2020). Difference-in-differences with multiple time periods.

De Grove, F., Boghe, K., & De Marez, L. (2020). (what) can journalism studies learn from supervised machine learning? *Journalism Studies*, *21*(7), 912–927.

Downie Jr, L., & Schudson, M. (2009). *The reconstruction of american journalism* (Vol. 20). Columbia University Graduate School of Journalism New York.

Fort, K. (2016). *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. John Wiley & Sons.

Friedland, L., Napoli, P., Ognyanova, K., Weil, C., & Wilson III, E. J. (2012). Review of the literature regarding critical information needs of the american public. *Unpublished manuscript submitted to the Federal Communications Commission. Retrieved from http://transition. fcc. gov/bureaus/ocbo/Final_ Literature_ Review. pdf*.

Gao, P., Lee, C., & Murphy, D. (2019, February 12). *Financing dies in darkness? the impact of newspaper closures on public finance* (SSRN Scholarly Paper ID 3175555). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3175555

Grant, N. (2020). The demographics of local news deserts within the regions of the united states.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages.

Hellekant Rowe, E. (2016). The state of local journalism: A comparative study of local journalistic output in three swedish municipalities.

Institutet för Mediestudier, b. (2021). Kommundatabasen [Accessed: 2021-04-06].

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kaplan, M., & Hale, M. (2010). Local tv news in the los angeles media market: Are stations serving the public interest? *Los Angeles, CA*.

Kekezi, O., & Mellander, C. (2017). *Geography and media: Does a local editorial office increase the consumption of local news?* Royal Institute of Technology, CESIS.

Krisinformation. (2018). *Msb issues an announcement to the general public about the extreme risk of fires*. https://www.krisinformation.se/english/news-in-english/old-news-in-english/msb-issues-an-announcement-to-the-general-public-about-the-extreme-risk-of-fires

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Social science. computational social science. *Science (New York, NY)*, *323*(5915), 721–723.

Mahone, J., Wang, Q., Napoli, P., Weber, M., & McCollough, K. (2019). Who's producing local journalism? assessing journalistic output across different outlet types. *DeWitt Wallace Center for Media & Democracy Report. https://dewitt. sanford. duke. edu/wp-content/uploads/2019/08/Whos-Producing-Local-Journalism _ FINAL. pdf*.

Miller, J. (2018). News deserts: No news is bad news. *Urban policy*, 59–76.

MPRT. (2019). Vägledning inför ansökan om mediestöd.

Napoli, P. M., Weber, M., MCcollough, K., & Wang, Q. (2018). Assessing local journalism. *New Measures Research Project. Available at: https://dewitt. sanford. duke. edu/wp-content/uploads/2018/08/Assessing-Local-Journalism_100-Communities. pdf*.

Nwanganga, F., & Chapple, M. (2020). Practical machine learning in r. *Practical Machine Learning in R*.

Nygren, G. (2020a). Challenges for local journalism in sweden. *The Routledge Companion to Local Media and Journalism*.

Nygren, G. (2020b). Coronaåret 2020.

Nygren, G., & Althén, K. (2014). *Landsbygd i medieskugga: Nedmonteringen av den lokala journalistiken och bilden av landsbygden i dagens nyheter*. Södertörns högskola.

Nygren, G., Leckner, S., & Tenor, C. (2018). Hyperlocals and legacy media: Media ecologies in transition. *Nordicom Review*, *39*(1), 33–49.

Nygren, G., & Nord, K. O. (2020). Svenska nyhetsredaktioner 2019: Krympande redaktioner – nätverksproduktion och läsardata i centrum.

Nygren, G., & Schjærff Engelbrecht, N. (2018). Vita fläckar och medieskuggor-stora skillnader i mediernas kommunbevakning.

Ohlsson, J. (2017). Medielandskapets dataficering.

Pålsson, P. (2017). 20 år av lokaljournalistik i strömsund.

Ramos, J. et al. Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*. *242*. (1). Citeseer. 2003, 29–48.

Retriever. (2021). Mediearkivet. https://www.retrievergroup.com/sv/

RKA. (2021). Kolada. https://www.kolada.se/

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Ryan, A. M., Burgess Jr, J. F., & Dimick, J. B. (2015). Why we should not be indifferent to specification choices for difference-in-differences. *Health services research*, *50*(4), 1211–1235.

Stites, T. (2011). Layoffs and cutbacks lead to a new world of news deserts. *Nieman Lab*, *8*.

Stoltz, D. S., & Taylor, M. A. (2019). Concept mover's distance. *Journal of Computational Social Science*, *2*(2), 293–313.

Strumpf, E. C., Harper, S., Kaufman, J. S., & Oakes, M. (2017). Fixed effects and difference in differences. *Methods in social epidemiology*, *1*.

Tenor, C. (2020). Gränser i förändring? journalistik och kommunal information i det lokala medielandskapet.

Truedson, L., & Karlsson, M. (2018). Nyheter – allt mer en tolkningsfråga? mediestudiers innehållsanalys 2007-2018.

Waldman, S. (2011). *Information needs of communities: The changing media landscape in a broadband age*. Diane Publishing.

Wallentin, L. (2020). Paus i redaktionsslakten.

Weber, M, Andringa, P, & Napoli, P. (2019). Local news on facebook: Assessing the critical information needs served through fakebook's todayin feature.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., & Attenberg, J. Feature hashing for large scale multitask learning. In: *Proceedings of the 26th annual international conference on machine learning*. 2009, 1113–1120.

Wihbey, J. P. (2019). *The social fact: News and knowledge in a networked world*. Mit Press.