CATÓLICA LISBON
BUSINESS & ECONOMICS

We are watching you:
Citizen journalism and accountability

Jonas Lüder

152119039

The dissertation is written under the supervision of
Prof. Dr. Christian Peukert
(Associate Professor of Strategy at the HEC Lausanne & Research fellow at the ETH Zurich)

Dissertation submitted for the
MSc in Strategy and Entrepreneurship,
at Católica Lisbon School of Business & Economics,
6th of January 2021

Abstract

Title:	We are watching you: Citizen journalism and accountability
Keywords:	Citizen journalism, corruption, accountability, corporate misconduct, readership attention
Author:	Jonas Lüder

Digital media reduces costs but also increases competition. Nowadays, costly reporting and the coverage of niche news are no longer cost-effective for publishing houses. I formulate hypotheses to test for reporting differences between citizen journalists and professional journalists. Specifically, this work aims to investigate for the intergroup differences of journalists when reporting topics of corporate misconduct. Differences in covering corporate misconduct seem to be not significant. However, with a sample of 665,507 observations, I state that citizen journalists do fill the gap by publishing articles complementary in topic to professional journalistic posts. The readership is significantly more attracted by news reporting containing names of companies that are published by citizen journalists. Furthermore, I control for article parameters such as length, pictures and comments which according to literature evoke more readership attention. The theory this research elaborates is to hold.

By operationalizing and controlling articles for keywords expressing corruption and for selected company names, I conclude that their use significantly increases readership attention. Based on the following analysis, I suggest that (1) digital newspapers should more actively include the public sphere in their portfolio by offering social platforms that enable participation, which in turn is incentivized by a rating index. In addition, (2) large corporates should include citizens as a sounding board for efficient communication due to their proximity to the public.

Sumário

Título da dissertação:	Estamos de olho em si: Jornalismo cidadão e responsabilização
Palavraschave:	Jornalismo cidadão, corrupção, responsabilização, má conduta corporativa, atenção dos leitores
Autor:	Jonas Lüder

Os meios digitais reduzem os custos mas também aumentam a concorrência. Hoje em dia, as reportagens dispendiosas e a cobertura de notícias de nicho já não são rentáveis para as editoras. Formulo hipóteses para testar as diferenças de reportagem entre jornalistas cidadãos e jornalistas profissionais. Especificamente, este trabalho visa investigar as diferenças intergrupais dos jornalistas quando noticiam tópicos de má conduta corporativa. As diferenças na cobertura da má conduta corporativa parecem não ser significativas. No entanto, com uma amostra de 665.507 observações, declaro que os jornalistas cidadãos preenchem a lacuna, publicando artigos complementares em tópicos a cargos jornalísticos profissionais. O público leitor é significativamente mais atraído pela reportagem de notícias contendo nomes de empresas que são publicadas por jornalistas cidadãos. Além disso, controlo os parâmetros dos artigos, tais como comprimento, imagens e comentários que, de acordo com a literatura, evocam mais atenção do público leitor. A teoria que esta pesquisa elabora é suportada pela analise dos dados coletados.
Ao operacionalizar e controlar artigos para palavras-chave que expressam corrupção e para nomes de empresas seleccionadas, concluo que a sua utilização aumenta significativamente a atenção do público leitor. Com base na análise seguinte, sugiro que (1) os jornais digitais deveriam incluir mais activamente a esfera pública na sua carteira, oferecendo plataformas sociais que permitam a participação, que por sua vez é incentivada por um índice de classificação. Além disso, (2) as grandes empresas deveriam incluir os cidadãos como caixa de ressonância para uma comunicação eficiente devido à sua proximidade com o público.

Acknowledgements

I would like to thank my teacher and dissertation supervisor Christian Peukert, who brought me many perspectives on the topic of digitization. Furthermore, I am thankful for his constant willingness to help me and his guidance with the right impulses to help me order my thoughts.

I would like to thank Moritz for his authentic feedback and being such a good friend.

To my mother for being such a strong role model and teaching me the right values and importance in life.

To Natalie, a true friend and discussion partner who makes my every day much more valuable.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

A growing literature exemplifies that search, transportation and reproduction decreased tremendously due to digital technologies (Goldfarb & Tucker, 2019; Shapiro & Varian, 1999; Waldfogel, 2017). This accelerates the exchange of information leading to constantly higher competition paired with decreasing switching costs for consumers. In consequence, newspapers find a strong transformation and have to adapt an efficient cost structure (Fogarty, 2012; Franck, 2019; McManus, 1994; Shapiro & Varian, 1999). Complex reporting and fragmented local news coverage must give way to strategies aimed at attracting the attention of the masses. Attention, respectively monetisation by clicks and views becomes the target currency for publishers (Franck, 2019; McManus, 1994). However, low costs and almost unlimited space for publications on the other hand provides opportunities of implementing product offers at a lower risk (Anderson, 2008; Peukert, 2019). Offering the public sphere a platform to participate in the journalistic cosmos possibly captures both, news coverage in categories "not worth covering" and consequently an expansion in online presence. Traditional publishing houses increasingly engage in this topic as scalability promises advertisement revenues (Goode, 2009). Citizen journalism has evolved as a new form of journalism, driven by a gap in local reporting activities and the urge of citizens to communicate. They have established themselves in the environment of news reporting, investigations in public affairs, business corruption and local events (Robinson & DeShano, 2011). The benefit of this "new form of journalism" is not only to be viewed from the platforms' perspectives, but also from the readers demand side. An echo of literature exemplifies that active media and a participatory society results in better governance by shedding light on moral misconduct (Enikolopov et al., 2018; Snyder & Strömberg, 2010).

Citizen journalism has greatly developed over the years and has become an integral part of everyday reporting and an investigative pillar of democracy (Robinson & DeShano, 2011). This assumption drives this work and leads my research to the empirical question of *how news coverage of corporate misconduct differs between professional and citizen journalists and how readership accounts for it*. My results support this hypothesis with having found significant differences in the topics these author types cover. Furthermore, I could find that citizen journalists tend to cover more topics with corporate reference. Additionally, citizen journalists attract significant more attention when covering corporate topics, which one can base on the theory of Reimers and Waldfogel (2020), who provide a possible explanation in the form of reader identification with the authors. One cannot state a significant difference in the reporting

of corporate misconduct between citizen journalists and professional journalists and further differences in the level of readership attention. The results support my hypotheses therefore only in parts. However, the models of this paper support clear recommendations for digital publishing houses to open their platforms for the public sphere. In a first step, the possibility to comment articles increases attention significantly and secondly articles posted by citizen journalists strengthen the network by covering local news in a cost-efficient way. Besides, citizen journalists' articles can be nearly as successfully as professional content with regards to readership attention. Intrinsic motivation and thus participation can or should be catalyzed by a weighted rating system analog to common online shops. On the one hand, such a rating motivates citizen journalist and on the other hand helps the readership to rate the author for its accountability.

A large body of literature exists about economics of mass media; however, the phenomenon of citizen journalism has rarely been investigated holistically with regards to supply and demand. In this paper I aim to fill this gap in a narrow sense and develop hypotheses regarding the role of citizen journalists and their accountability. Moreover, I investigate for characteristics of an article defining the potential success in terms of readership attention, while supporting the statement about the importance of text length, use of picture and keywords of Andrews and Caren (2010).

The following work is introduced by a literature review about journalism, mass media and the role of citizen journalists. That is followed by an overview about the data set I use, the methodology and model applied in the following research. Last, I present my analysis and discuss the results of my models.

2 Literature Review

The following chapter orchestrates a relevant literature review of interpretive journalism research, mass media and the role of citizen journalism as an accountable local reporting mechanism. Furthermore, I aim to provide a background on corporate misconduct and consequently formulate my hypotheses.

2.1 Interpretive journalism and the value of mass media

Before answering my research question, in this section I aim to provide a background on the role of journalists in the sphere of digital news coverage. To better understand the occurrence of citizen journalism, it is essential to understand the mechanism of digital mass media.

Interpretive journalism

The interpretive media landscape was dominated by a professional journalistic system collectively united by shared standard values that do create a journalistic interpreting and reporting authority (Berkowitz & TerKeurst, 1999; Kovach & Rosenstiel, 2014; Zelizer, 1993). Journalistic paradigms in this cosmos are formed by internal and external contacts, its status, knowledge and capabilities, that manifest journalistic reporting authority towards the general public (Zelizer, 1993).

By definition according to Patterson (1996), interpretive journalism is the exact opposite of descriptive journalism. The definition of interpretive journalism goes beyond replicating information by contextualising and enriching the facts with deep analysis and conclusions of possible implications (Keller 1997). High levels of interpretation in journalistic articles make journalists rather to "makers of meaning" than to passive news transmitters (Barnhurst, 2013). Contrary to Patterson (1996), the theory of Agee et al. (1983) sees interpretation inevitably linked to journalistic work. However, there is a consensus in theory that interpretive journalism creates additional meaning by holistic context reporting and assessment of the facts. Contextualized reporting becomes increasingly important in more complex contexts and should therefore empower citizens to grasp interrelations by critical journalistic news reporting (Byström, 2002; Krieghbaum, 1956; Patterson, 1996; Schudson, 1996). Monopolistic news coverage by journalists is viewed critically, since journalists as gatekeepers potentially further divide the distance between citizens and politics and its key actors (Goode, 2009; Patterson, 1996). Further, new journalistic participants immigrate into the journalistic system and might adopt to dominant norms (Goode, 2009).

Digital mass media

What people cannot experience directly personally, they do over a medium, that provides information for interpretation (Pennington, 2012). News media is thus having a pivotal role in shaping the public agenda by navigating the discussion on events and issues (Andrews & Caren, 2010; Edwards & Wood, 1999). Journalists as transmitting navigators build reports out of pre-

sorted information. This function also leads to the so called definition of journalists as "gate-keepers" (Andrews & Caren, 2010; Goode, 2009; Patterson, 1996; Shoemaker & Vos, 2009). Fundamental to this news reporting process, is subjective interpretative journalism. Literature exemplifies the consequence of a journalistic media bias for the public, which can be observed in form of selection, description or reliability and validity bias during the gate-keeping process of journalists (Franzosi, 1987; McCarthy et al., 1996).

In the digital age, the need of selection for publishing houses is driven by limited "news holes" in form of space and sources available that allow easy access to enable quick publications (Andrews & Caren, 2010; McCarthy et al., 1996, p. 480). Convenience and accountability in mainstream sources potentially result in outcome sets with neglection of information (Fogarty, 2012; Gamson et al., 1992; McCarthy et al., 1996; Shapiro & Varian, 1999).
However, positive effects of digital technologies in news media are remarkable. Search, transportation, reproduction and validation cost decrease, while returns increase due to digitization of information technologies.
Nevertheless, in the information market, lower costs are leading to a higher degree of competition intensity with low consumer switching costs (Goldfarb & Tucker, 2019; Shapiro & Varian, 1999; Waldfogel, 2017). As a consequence, only information services respectively publishing houses who adapt their cost and revenue structure to a global world do survive (Shapiro & Varian, 1999). This ultimately leads to a consolidation of the information market. Business units such as local news coverage shrink down due to high resource utilization, budget cutbacks and markable space (Fogarty, 2012; Franck, 2019; McManus, 1994). Additionally, fragmented and unofficial sources are costly to cover and hard to scale on a national level (ibid.). Hence, a global flood of overlapping convenient news coverages reaches the readership.

Research attempts that the gatekeeper role of journalists is challenged for financial purpose in the cosmos of digital mass media arguing that the flood of lower quality and convenient journalism is at the costs of the civic readership (Franck, 2019; Shapiro & Varian, 1999; Waldfogel, 2017). Furthermore, negative impacts on social stability and public accountability due to a lack of local news coverage occur (Brunetti & Weder, 2003; Campante & Do, 2014; Prat & Strömberg, 2013; Snyder & Strömberg, 2010).
However, one must also keep in mind that the digital media is also to the consumers' advantage. The decrease in inventory and search costs brings a rich variety to the civic society (Brynjolfsson et al., 2003; Chiou & Tucker, 2017).

Increased speed of information distribution and cheap information access results in an information overload for the demand side, ultimately leading to a lack of readership attention (Shapiro & Varian, 1999). Consequently, news coverage changes to an economy that is driven by readership attention measured with views, clicks, likes etc. that can be monetised by advertisement placements (Franck, 2019; McManus, 1994). In order to enhance attention, articles need to include specific features such as famous, impactful and spectacular (notorious, consequential and extraordinary) keywords (Andrews & Caren, 2010; Franck, 2019; McCarthy et al., 1996). Further, the nature of the articles (length, graphics, structure) does play decisive role in the race for the readership attention (Andrews & Caren, 2010; Koopmans, 2004; Rayner et al., 2008).
Due to the theoretical elaboration of the previously mentioned authors, I expect that attention increases with the use of keywords, ceteris paribus and my hypothesis to hold true:

Hypothesis 1 (H1) *The use of keywords increases the readership attention.*

2.2 Citizen Journalism

Before discussing the phenomena of citizen journalisms, a clear definition around the terminology must be drawn. *Citizen journalists are independent amateur users who practice journalistic activities on community affairs at collegial news sites* (Robinson & DeShano, 2011).

This section focusses on the relevance and diverse body of participatory, network or grassroot civic journalism, also defined as citizen journalism (Bowman & Willis, 2003; Gillmor, 2004; K. A. Johnson & Wiedenbeck, 2009). Citizen journalists fundamentally challenge traditional gatekeeping roles in news coverage by turning the journalistic top down navigation to a horizontal conversation between the public sphere, press and institutions (Gillmor, 2004; Goode, 2009; Habermas, 1999). The change in the cost structure and an easy access to sources has greatly facilitated the creation of new articles by transforming the civic population from consumers to active players in the media (Andrews & Caren, 2010; Waldfogel, 2017).
The public sphere becomes citizen journalists that report, (re)-interpret and investigate outside the mainstream media (Atton, 2015; Bolter & Grusin, 1999; Gillmor, 2004). Citizen journalism is generally executed on web based platforms allowing interactions between news publishers and readers via comments and direct reactions, which dilutes the exact definition of citizen

journalism activities (Goode, 2009). Besides a variety of definitions, my research defines citizen journalists as non-professional journalistic direct creation of content (Goode, 2009; Kaufhold et al., 2010). They are participating on dedicated social news sites or their content directly emerges into reporting value chains of newspapers websites (Gillmor, 2004; Goode, 2009). Quite often they are their own source and set the agenda by reporting and interpreting topics they have concerns about (Fogarty, 2012; McCombs, 2004).

The symbiosis of professional and citizen journalisms provides for complex interrelations leading to a wider news coverage offer to the readership. Consequently, citizen journalists fill the local news hole publishers preferably do not cover due to financial stress from global competition. Citizen journalists cover reporting niches and form the "long-tail" on news platforms, subsequently let publishers test more with lower risks (Anderson, 2008; Peukert, 2019; Waldfogel, 2017, p. 201). Offering a broad news portfolio, publishers strengthen their online network and attract more users, by the increased probability of random reporting successes in the "long-tail" (Anderson, 2008; Chiou & Tucker, 2017; George & Hogendorn, 2013; Peukert, 2019). In consequence, more data is available, which allows for better predictions and readership targeting for digital mass media publishers (Peukert & Reimers, 2018). It is of interest, how citizen journalists adapt to the above outlined journalistic system, which constantly gains in complexity. I conclude that citizen journalists do cover different news categories in comparison to professional colleagues, ceteris paribus:

> **Hypothesis 2 (H2)** *Professional journalists and citizen journalist do differ in the news categories they cover.*

News coverage of corporate misconduct

Pointing out the relevance of citizen journalists in the online media sphere, citizen journalists may adapt norms of mainstream journalism when interacting. I am interested in their contribution to critical news coverage and the way the readership responds. Following, I describe the term of accountability and reporting of corporate misconduct. Subsequently, I derive my final hypothesis distinguished between the journalistic groups to answer my research question of *how news coverage of corporate misconduct differs between professional and citizen journalists and how readership accounts for it*.

Accountability

Literature observes growing concerns about falling trust of society in professional news media due to persuasion biases of the factual journalistic accountability (Campante & Do, 2014; Gentzkow & Shapiro, 2006). Also, online reviews can mislead consumers in terms of perceived accountability (Mayzlin et al., 2014). The accountability entitled to citizen journalists is therefore called into question, as prestige sources like interpreting professional journalists still lead to a higher credibility on the readers' site according to literature (Hovland & Weiss, 1951; K. A. Johnson & Wiedenbeck, 2009). In parallel, the proportion of citizen journalism news coverage in the digital media is increasing in influence (Carr et al., 2014).

The readership faces many different sources from various author types, which are hard to account for quickly. Web accountability of citizen journalists can be increased by providing a journalists profile with sources, credentials and ratings (K. A. Johnson & Wiedenbeck, 2009). Social proof helps the consumers to evaluate the quality of the articles (Mayzlin et al., 2014; Reimers & Waldfogel, 2020). The more reputation an active journalist has, the higher the perceived accountability by the readership and its engagement when reading the article (K. A. Johnson & Wiedenbeck, 2009).

Corporate misconduct

Journalism and mass media are of great significance in democracies. Active media can be a corruption limiting filter element leading to better governance (Enikolopov et al., 2018; Snyder & Strömberg, 2010).

Citizen journalists become an important role in corruption investigations and raising massive public attention on corporate misconducts (Healy & Ramanna, 2013). This leads us to further investigate the online news coverage of citizen journalists in terms of corporate misconduct.

I define corporate misconduct for further research as corruption in a business context.

Since corruption is a social phenomenon that emerges in many fields and facets, a precise definition is difficult to draw (R. A. Johnson & Sharma, 2004). However, for the purpose of this research paper corruption can be defined as *"private gain from public office",* translated into seeking a rent out of bribe (Brunetti & Weder, 2003; Gardiner, 2017; R. A. Johnson & Sharma, 2004; Stark, 1997, p. 108).

In order to specify the construct of corruption in this context, I operationalized the definition of corruption with the associated keyword list of R. A. Johnson & Sharma (2004) Table 7.

Professional journalists possess more knowledge due to their resources and capabilities. However, competition in news production limit the amounts of costly and time intensive journalism in the race for fast readership attention. Ceteris paribus, this suggest that citizen journalist tend to cover news in form of corporate misconduct differently. I derive the following hypothesis:

> **Hypothesis 3 (H3)** *Citizen journalists are more likely to cover corporate misconduct than professional journalists.*

According to Hovland & Weiss, (1951) and Johnson & Wiedenbeck, (2009) professional journalists are assigned with a higher level of accountability. I hypothesize, that readership pays more attention to professional journalists covering corporate misconduct, ceteris paribus:

> **Hypothesis 4 (H4)** *The effect on readership attention of covering corporate misconduct differs between professional journalists and citizen journalists.*

H4 seeks to examine the combination of accountability, the use of corruption keywords and the type of author.

3 Austria and Corporate Misconduct

Misconduct in a business context is more precisely defined as corporate corruption for my purposes. The concept of corruption is not necessarily the same across countries and thus bothers the public sphere differently (Gardiner, 2017). In this paper, I testified my hypotheses on the example of Austria, which places number 12 (1 best) according to the corruption perceptions index (Transparency International, 2019).
Austria has a rich journalistic infrastructure and a broad political public participation, possibly taking up corruption affairs via new mass media (Prat & Strömberg, 2013; Snyder & Strömberg, 2010; Mohr, 2020).
This makes Austria an ideal example for investigating the supply and demand side of articles about corporate misconduct by testing my hypotheses.

Corruption in Austria

Austria represents a western democracy, nevertheless corruption does play a role in the context of corporate business. Taking the corruption perceptions index (CPI) into consideration, Austria

significantly improved by eight CPI points from 2012-2019. However, the paper is motivated by the recent overall Austrian CPI rating of 2019, in which Austria places twelve and is in general mostly affected by corruption in the business sector (Riano et al., 2010). This results in considerable financial damage in Austria every year, which is difficult to quantify fully due to disclosure in the mostly advantageous corruption activities for parties involved.

The website "meinbezirk.at"

The following provides a short overview about the Austrian online news site "meinbezirk.at", which serves as a foundation for my research. Meinbezirk.at is an equal joint venture of the Moser Holding Aktiengesellschaft and Regional Medien Austria (RMA) Beteiligungsverwaltungs GmbH, initiated in 2009. The RMA is built on three pillars: local, regional and national news coverage in Austria. In total, RMA issues 126 local and 121 online newspapers by 88 local offices in Austria (Regionalmedien Austria, 2020).

The entire RMA portfolio is summarized online by the site meinbezirk.at. In addition to employed editors, paid freelancers, institutional publications and commercial journalists, the platform is open to citizens to publish their articles with a private account. So called "Regionauts" form the unpaid long-tail as participating citizens, which simultaneously expand the network with niche content and a broader geographical coverage. The type of the author is displayed in the author profile to the readers, which is supplemented by a point index, followers, reach and accession date. The point index rewarding participation has been removed in 2018-09. This is however not due to a strategic decision but rather for a cosmetic layout reason.

News are published in chronological order under the authors profile. Citizen authors receive direct feedback on their publications in a personal performance dashboard and can benchmark the success of their articles. A selection of online published citizen articles is released in the RMA newspapers. In Austria, meinbezirk.at achieves a monthly penetration market rate of 33.6% and a growth rate of 6% year on year 2019 (Regionalmedien Austria, 2020; Meinbezirk, 2020; OEWA Plus, 2020).

Specific characteristics of "meinbezirk.at"

Meinbezirk.at summarizes a variety of author types, which may be beneficial for meinbezirk.at and in consequence to the shareholders. The inclusion of the community may lead to advantages in cost structure, network and readership attention in the competition in Austria's news media market. It is worth summarizing the article characteristics divided by the author types as their

differences base the focus of my research. In the following, I summarize the site-specific distribution:

Figure 1: Summary statistics by Citizen Journalists (=1)

	0 (N=451150)	1 (N=214357)	Total (N=665507)
#Views			
Mean (SD)	322 (1100)	211 (813)	286 (1020)
Median [Min, Max]	111 [0, 134000]	100 [0, 131000]	107 [0, 134000]
#Words in Title			
Mean (SD)	6.08 (2.32)	5.79 (2.92)	5.99 (2.54)
Median [Min, Max]	6.00 [1.00, 34.0]	5.00 [1.00, 43.0]	6.00 [1.00, 43.0]
#Words in Text			
Mean (SD)	180 (148)	167 (172)	176 (156)
Median [Min, Max]	142 [1.00, 4830]	123 [1.00, 8260]	136 [1.00, 8260]
#Pictures			
Mean (SD)	7.89 (26.5)	8.39 (23.6)	8.05 (25.6)
Median [Min, Max]	1.00 [0, 1310]	2.00 [0, 924]	1.00 [0, 1310]
#Comments			
Mean (SD)	0.0531 (0.507)	0.597 (2.88)	0.228 (1.71)
Median [Min, Max]	0 [0, 129]	0 [0, 84.0]	0 [0, 129]

4 Data and Methodology

The following section presents the data origin and the construction of the final data set on online news coverage and explains the econometric methodology applied for this research.

4.1 Data Source and Structure

To answer the question of accountability of citizen journalists, one requires insights on the performance and thematic contents of all participating journalists in order to be able to put them in relation. The source of my data set is the data of the Austrian website "meinbezirk.at". The dataset provided to me, is rich in information and gives a full summary of all journalistic activities of the years 2007-09 to 2019-02 on the website. Posted articles are collected chronologically on an hourly level assigned to its author and further descriptive metadata. The raw dataset summarizes 2,119,318 observations. I limit the time period to daily data from 2015-2019, since one does not have a tracking of clicks per article earlier than 2015 (891,986 observations). Further I drop observations, not stating the region mostly paired with missing information in the type of author (18,347 observations), data missing in variables (24,574 observations), observation with no word in text or abstract and no pictures (52 observations), more general posted article without text (156,176 observations), articles not posted by an individual but functional author (103,946 observations) and articles which are an event post (157,103 observations). I further limited the categories of news coverage to the most relevant for my research leading to a final data set of 665,507 observations summarizing the unique article observations of my sample from meinbezirk.at.

For further analysis, I operationalise the authors' text and use looping text mining algorithms in R Studio to retrieve patterns of the published articles and make corporate misconduct measurable. Using the posted article as input information and treating each string as a single object. Text mining transforms the published article into a structured format, which allows me to discover frequencies (Feinerer et al., 2008). In a first step, I adapt the letters to lower cases and adjust for special characters in the substring (article) and my search string (keywords). I build a counting continuous variable for every string, with a summarizing binary response variable building the coefficient *Corporate Misconduct.*

The variables in my data set are described in Table 8 and are explained in more detail below in chapter 5 . As data source for the corporate search string I refer to the "Vienna Stock Exchange"

and the newspaper “Trend.at”, that are summarizing most important companies listed at the Austrian stock market or being important in size.

4.2 Selected Types of Regression Models

The journalistic focus of this research in the first step suggests a qualitative investigation. However, the data set provided allows to deal with the topic quantitatively and does represent a significant sample due to its size and market share in the Austrian news media industry.

Answering my hypotheses requires a multivariate approach, due to the continuous and binary nature of my dependent variables. The investigation of demand in form of readership attention in H1 and H4 is based on the dependent continuous variable *views,* that allows estimates applying multiple linear regression. I apply the logarithm to my output variable, meaning that the log of *views* tends to follow to a normal distribution as some article posts signal outliers in higher readership attention. Empirically, I include all control variables to have robust results and address potential endogeneity in the model that correlate with my input factors and explain my output. Thus, I remove any time constant in the error term $u(i)$ by applying calendar-year fixed effects to my model. Fixed effects limit selection bias and eliminate variations in-between units by averaging variables effect (Mummolo & Peterson, 2018; Wooldridge, 2016). $\varepsilon(i,t)$ remains as element of the error term including the time variant errors. Additionally, the multiple linear regression allows us to test for other factors by including control variables potentially correlating with my input variables. To avoid multicollinearity, I consider m (variable specifications)-1 when including category dummy variables in my regression model.

The holistic answer to my four hypotheses requires in addition to the multiple linear regression a more detailed review of the second *output* variable *Citizen Journalist*.
The *output* of H2 and H3 is a binary depend variable and consequently a probit or logit model, due to the nature of the variable *Citizen Journalist.* However, I use a linear probability model (LPM). I include as further input variables (category), state and calendar-year fixed effects. I include the dummy variable *category* not in the fixed effects as I am moreover interested in these coefficients in my regression output. Whereas the specific characteristics of interaction terms using discrete choice models are addressed by Mummolo and Peterson (2018), multiway fixed effects theory is lacking in explanation (Claussen et al., 2018). Year and state fixed effects in my model with the output variable *Citizen Journalist* do describe a multiway fixed effect and are inevitably for exogeneity in my regression. Additional the inclusion of interaction terms

lead me to follow the approach of Claussen et al. (2018). I pass by the pitfall of a skewed distribution in my *output* variable which is turning to an estimation bias when using panel data. However, I apply the probit model in my econometric analysis for the *Citizen Journalist* variable in appendix Table 5.
Further, I examine the hypotheses presented in chapter 2 with model specifications well-based in research presented hereafter.

5 Specific Regression Models and Variables

In this chapter, I present the variables which will codify my regression models in more detail. On the one hand I deal with the question of readership demand in H1 and H4, on the other hand I investigate supply and characteristic of articles in H2 and H3. For this purpose, the analysis is based on two theoretical foundations, which are outlined below with the associated dependent variables and an including model specifications. I start with the description of the independent variable in need of further explanation in addition to the variable explanation Table 8.

5.1 Independent Variables

Index Removal

The term of the index removal previously displayed besides the article is treated as a shock for the supply side in this paper. An index can be perceived as a reciprocity incentive for unpaid participation in reporting (Cammaerts, 2011). A shock for the demand side would also be conceivable, because as described in chapter 2.3, an index can be beneficial for accountability and consequently for attention. But the present index does not describe the quality of the author and article, as the point system of the index is linked to participation quantity in the social network. The variable *Index Removal* summarizes binary the pre and post period of the availability an author index, where one states the post-index period (*Index Removal)*.

5.2 Dependent Variables

Attention *(Views)*

As a theoretical foundation for the investigation of H1 and H4, which sheds light on the demand side, Marshallian demand theory (Coto-Millán, 2003) is used. I assume that different input

factors contribute to a change in demand and consequently in readership attention (Baye & Prince, 2014).

I am interested in the effect of keywords use on attention and further on the difference between different authors covering corporate misconduct. In a more condensed model for my research foundation, I apply the model of Chiou & Tucker (2013) in which they investigate for visits to website in dependency of input factors. In addition, Reimers & Waldfogel (2020) do set up an information utility function, which I slightly modify. R_i is the rating on product i and x_i other observables, I model:

$$views_i = v(R_i; x_i) \tag{1}$$

The model serves me as a foundation for the inclusion of the social index variable. To explore the interrelation in a formal way, I specify the model in reference to the empiricism of Chiou & Tucker (2013) and formulate the following regression for article *i* in year *t* and category *c*:

(2)

$$\begin{aligned} &\boldsymbol{log(views_i)} \\ &= \beta_0 + \beta_1 CitizenJournalist \times IndexRemoval_t \\ &+ \beta Keywords_{ic} + \beta Controls + \beta FE_{itc} + \epsilon_{itc} \end{aligned}$$

Readership attention (*views*) forms my *Output* variable. The coefficient βFE_{it}[1] summarizes the article category fixed effects, state fixed effects and the year fixed effects.

As pointed out in chapter 2.1 we expect that the level of readership attention is to increase dependent on input coefficients characteristics. I am mainly interested in the effect of *Keywords* and *Citizen Journalist*. Thus, I include all measurable factors and use fixed effects in order to isolate the effect and reduce endogeneity. Equation (3) describes the interaction term β_1 in detail included in regression (2).

(3)

$$\frac{\{\mathrm{E}|views|CitizenJournalist = 1,\ IndexRemoval = 1|\}}{\{\mathrm{E}|views|CitizenJournalist = 1,\ IndexRemoval = 0|\}} = exp(\beta_1)$$

[1] FE = Fixed Effects

Corporate Misconduct

Following, I focus on the supply side and hypothesize in H2 and H3 differences in the news coverage in general and consequently more specific of corporate misconduct between professional and citizen journalists.

The previously formulated definition in chapter 2.3 and the text mining method described in chapter 4.2 enables me to form a measurable coefficient for *Corporate Misconduct.* The coefficient is built on the occurrence of company names and the operationalized definition of corruption Table 7.

Due to the binary nature of the dependent variable, I base the analysis of H2 and H3 on a different theoretical foundation than H1 and H4 above. Brunetti and Weder (2003) analyze corruption in dependence of the level of press freedom and form their dependent variable *Corruption* to an integer variable. Consequently, they use a discrete choice model, which serves us as reference in the type of *output*. I specify my model for the dependent variable of *citizen journalists* in which one states that the journalists is a citizen and zero a professional journalist:

(4)

$$\boldsymbol{CitizenJournalist} = \beta_0 + \beta_1 Corp.Misconduct \times IndexRemoval + \beta Category + \beta Controls + \beta FE + \varepsilon$$

I am mainly interested in the interpretation of the coefficient β_1 and β_2 in order to test and answer H2 and H3. The variable *Category* has variants that I am testing in my regression. Analog to (2) I include control variables and βFE[2] for fixed effects including calendar-year and state. The control for fixed effects includes in contrast to model (2) different challenges in estimation using *Citizen Journalists* as dependent variable, previously specified in chapter 4.2. Equation (5) describes the interaction term β_1 in detail included in regression (4).

(5)

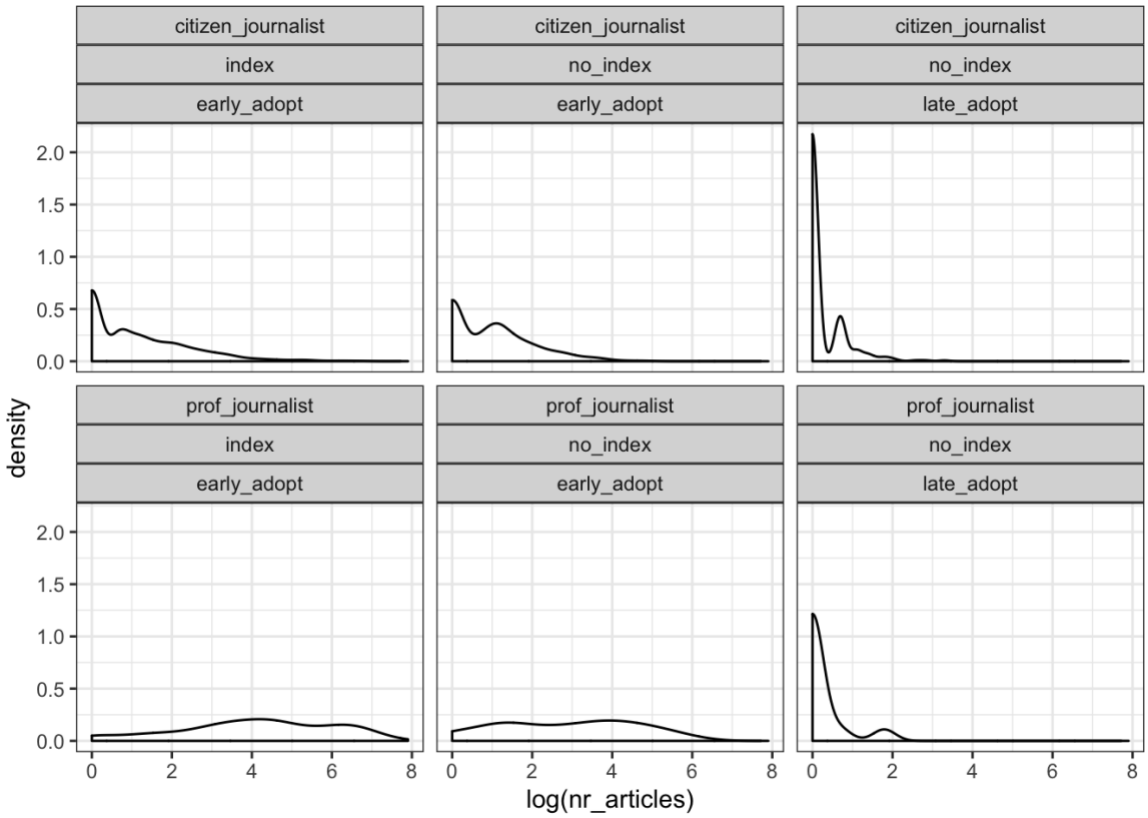
$$\frac{\{\mathrm{E}|CitizenJournalist|Corp.Misconduct = 1,\ IndexRemoval = 1|\}}{\{\mathrm{E}|CitizenJournalist|Corp.Misconduct = 1,\ IndexRemoval = 0|\}} = exp(\beta_1)$$

$$\{\mathrm{E}|CitizenJournalist|Corp.Misconduct = 0,\ IndexRemoval = 1|\}\ \{\mathrm{E}|CitizenJournalist|Corp.Misconduct = 0,\ IndexRemoval = 0|\}$$

[2] FE = Fixed Effects

The following graph shows the change of supply for the author groups *Citizen Journalists* and *Professional Journalists,* that started to publish before and after the index was removed and the pre and post time of the index removal (exact numbers summarized in Figure 3).

Figure 2: Impact on journalists of the index removal in 2018



Controls

The reason why my dependent variables can be affected by input factors might be due to many different reasons, which have to be controlled within my possibilities of data available. Besides the earlier mentioned fixed effects, I control for characteristics of published articles, that either can be relevant for the attention of the readership or be specific to the type of author.

Calendar-years are summarized in the fixed effects in order to control for characteristics that are related to the time an article is published online. Also, I do control for *state* as differences driven by demographics and geographics of authors' locations that need to be controlled for. Additionally, I control for the category an article is classified under, in order to prevent endogeneity but more from a research perspective, as I expect to see differences in the general news coverage comparing citizen journalists and professional journalists. Consequently, I control for the article meta characteristics, since I assume that they not only correlate with my

output but state a significant explanatory part of the input variables. I use the logarithm of the variables *words in text, number of pictures, words in title and no of comments*.
The expected impact on article demand is not clear, since readership attention can be increased with quantity and/or quality but also decreased. I expect the journalistic supply slightly to differ in length and complexity due to its functional focus (profession) and time capacities.

6 Analysis

In the following part I outline the results of my empirical analysis. The interest of my study focuses on the supply of news coverage of corporate misconduct and its readership reception. The persistent research echo about the different determinants impacting media attention (Andrews & Caren, 2010) is addressed by answering H1. This paper further enriches the research on citizen journalists with the answer to H2, H3 and H4. As mentioned in chapter 4.2, I run regressions separately for the *output* variables, namely *log(views)* and *citizen journalist*. For readability, I summarize the main regression in the following analysis part, whereas the regressions that leads me to the conclusion are attached in Table 3 and 4.

Table 1: Results for readership attention

Readership Attention log(Views)

	Dependent variable:			
	ln_no_views			
	(1)	(2)	(3)	(4)
citizen_journalist		-0.262***	-0.264***	-0.263***
		(0.003)	(0.003)	(0.003)
iscorruption	0.149***	0.173***	0.147***	
	(0.013)	(0.016)	(0.013)	
iscompany	0.011	0.013	-0.024	
	(0.024)	(0.024)	(0.029)	
corp_corr				-0.110
				(0.244)
iscorr_title	0.565***	0.545***	0.547***	0.600***
	(0.026)	(0.026)	(0.026)	(0.025)
log(words_in_text)	0.245***	0.236***	0.236***	0.237***
	(0.002)	(0.002)	(0.002)	(0.002)
ln_no_pictures	0.346***	0.352***	0.352***	0.352***
	(0.001)	(0.001)	(0.001)	(0.001)
log(words_in_title)	0.255***	0.246***	0.246***	0.246***
	(0.003)	(0.003)	(0.003)	(0.003)
ln_no_comments	0.311***	0.363***	0.364***	0.364***
	(0.004)	(0.004)	(0.004)	(0.004)
iscorruption:iscompany	-0.380*			
	(0.216)			
citizen_journalist:iscorruption		-0.087***		
		(0.029)		
citizen_journalist:iscompany			0.127**	
			(0.053)	
citizen_journalist:corp_corr				-0.502
				(0.498)
Article Category FE	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	665,507	665,507	665,507	665,507
Adjusted R2	0.258	0.267	0.267	0.267
Residual Std. Error	1.069	1.063	1.063	1.063
Note:		*p<0.1; **p<0.05; ***p<0.01		
	Table summarizes Table 1.1 and 1.2			

Results for control variables

I include article specific variables that potentially impact on attention Table 3. The coefficients shown in Table 1 column (1) *log(words_in_text), ln_no_pictures*[3]*, log(words_in_title)* and *ln_no_comments*[1] are positive correlated with *ln_no_views*[3]. I find evidence for the statements of Andrews and Caren (2010) at a 1% statistical significance level. The control coefficients are

[3] *no* = number of

of the expected sign, meaning, that if I change the control variables by one percent, I expect attention (*ln_no_views)* to increase by roughly rounded about 0.25%-0.35% ceteris paribus, stated in column (1). The time variant factor years is controlled for in the fixed effects in all models, further testing for article category in fixed effects improves my model with regards to the Adjusted R2, Table 3 and 4.

I observe that the characteristic attributes of my control variables following the structure of an article, does play a key role in addressing readership attention.

Results for MLR[4] hypotheses H1 and H4

Besides the article structure, media signals such as keywords do play a major role in attracting mass audience attention (Andrews & Caren, 2010). In Table 4 I run my regression to preliminary test for corporate misconduct column (1) and show the elasticity of an article covering *corporate corruption* by controlling for article characteristics. Adding the explanatory variables *iscorruption, iscorr_title, iscompany* and fixed effects for *state* and *year* improves the fit of my regression (adjusted R2 from 0.189-0.240). Interpretation of my log-transformed *output* variable with keywords in text or title dummy variables are to be interpreted in percentage change in *ln_no_views* relative to the omitted category. Looking at Table 1 column (6) *Log(views)* is approximately 16.1% (0.149) percentage higher for articles covering corruption respectively 76% (0.565) for articles having corruption keywords in its title, both at a 1% statistical significance level, holding other variables fixed. Thus, attention is higher for articles having corruption keywords in the title or text, holding other variables constant which supports H1. The negative sign of the coefficient company keywords is unexpected but is insignificant. Additionally, I add the interaction term of *iscorruption*iscompany,* which describes corporate misconduct when both variables state one. The interaction *iscorruption*iscompany* captures the extent of -46% (-0.380) to which *views* change at a statistical significance level of 10%. Meaning that articles about corporate misconduct (*iscorruption*iscompany)* do decrease the level of attention (*views*). Focussing on the article characteristics (*controls*) in Table 1 (1-4), H1 is largely confirmed.

Testing for H4, I include the variable *citizen journalist* in the model. Table 1 columns (2) to (4) summarize the regression of focus to test for readership attention in dependency of the news coverage of citizen journalists. Column (4) Table 1 presents the final estimate for readership attention. In Column (2) and (3) Table 1, I interact *citizen journalist* with *iscorruption*, which

[4] multiple linear regression

decreases readership attention by -8.3% *and iscompany* increases attention by 13.5%, both at a statistically significance level of 1%. Thus, I find first support for difference in demand of news coverage between professional journalists and citizen journalists. To finally test H4, I include the variable *corp_corr* in column (4) in order to be able to describe more specific *corporate misconduct* in interaction with *citizen journalist.* In difference to (2) and (3) the coefficient of citizen journalist covering corporate misconduct (*citizen_journalist*corp_corr)* does not show statistically significance at all. This demonstrates that readers do not significantly differ between citizen journalists and profession journalists when paying attention to articles covering corporate corruption. A negative, but insignificant estimator for attention can be seen. However, it cannot be found any support for H4. The effect on readership attention of covering corporate misconduct differs not between professional journalists and citizen journalists.
However, column (2) and (3) do show valuable findings on differences between authors and news coverage worth noting and will be further discussed in chapter 7.
Overall, the results presented in Table 1, state that there is a clear impact of characteristics and keywords used on readership attention. Consequently, my research provides additional support for the literature mentioned in chapter 2.1 and can further extend the knowledge within the research about citizen journalism in online mass media.

Results for LPM[5] hypotheses H2 and H3

Finally, I investigate H2 and H3 with the limited dependent variable *citizen journalist.* The regression in Table 2 represents a viable alternative to logit and probit models with the use of multiple fixed effects and allows reasonable interpretation (Angrist & Pischke, 2009; Claussen et al., 2018).

In column (1) Table 2, I test H2 and find evidence that supports for H2, as I control for probabilities of differences in the coverage of news categories. I find positive and negative relationships of news categories with *citizen journalist.* Hence, news categories do differ between the author groups. The stated statistical significance in my LPM provides me a guide to analyse the differences between professional journalists and citizen journalists, further tested and underlined within the probit model Table 5. I control for *state*, *year* and article characteristics namely summarized as *controls*.

[5] linear probability model

In column (2) Table 2, I add *iscorruption, iscompany and iscorr_title* in order to test for H3. Contrary to my preliminary expectations, I do not find any significant connection between *iscorruption* and *citizen journalist*. This suggest no difference in the corruption coverage within the author types. However, *iscompany* does state a significance. The positive relationship indicates that *citizen journalists* tend to cover more news having company keywords in the article compared to professional journalists. Contrary to the negligible effect of corruption mentioned in text, *iscorr_title* highlights a negative association with *citizen journalist*. Professional journalists tend to make more use of keywords within the article title than citizen journalists.

In column (3), I introduce the input factor *index removal* and find evidence for treating *index removal* as a shock for the supply side in Figure 1, as productivity decreases noticeably. Consequently, having a clear line between times where citizen journalists participate more productive than after the removal. The removing of the index has a negative association with the output variable *citizen journalist*, state that the index usage is to be an incentive for participation column (3). Taking this into consideration in column (4), (5) and (6), I interact *index removal with iscorruption, iscompany and iscorr_title.* The coefficient results remarkably change to column (2) and (3) in its coefficients. Important to highlight is the interaction term *iscorruption*index removal* in column (4), stating a negative association for covering corruption after the point index is removed with citizen journalists. In column (6) add the interaction term *iscorr_title*index removal* that also states a negative interaction term. Finally, in column (7) I address H3 and my research question of supply differences between authors covering corporate misconduct. I add the interaction term *corp_corr*index removal* and find a negative association between the interaction term and the probability that the article is published by a citizen journalist.
While I confirm H3 in column (2) to (6), I find no support for differences in corporate misconduct coverage.

Summarizing my results, the variable of interest shows no statistical significance on readership attention and further not for news coverage differences between professional journalists and citizen journalists. However, the results draw my attention on the factors that published articles contain topics about corruption and separately thereof do cover corporate topics. I discuss my results presented in my discussion in chapter 7 and relate my findings with the relevant literature I have reviewed above.

Table 2: Linear Probability Model “Citizen Journalist”

LPM Citizen Journalist

	Dependent variable:						
	citizen_journalist						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
factor(category)local	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
factor(category)news	-0.622***	-0.621***	-0.630***	-0.630***	-0.630***	-0.630***	-0.631***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
factor(category)people	0.017***	0.017***	0.017***	0.017***	0.017***	0.017***	0.016***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
factor(category)politics	0.041***	0.041***	0.041***	0.041***	0.041***	0.041***	0.041***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
factor(category)sports	0.080***	0.080***	0.080***	0.080***	0.080***	0.080***	0.080***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
iscorruption		0.001	0.0004	0.004	0.0004	0.0004	
		(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	
iscompany		0.027***	0.027***	0.027***	0.027***	0.027***	
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	
iscorr_title		-0.068***	-0.067***	-0.066***	-0.067***	-0.058***	-0.067***
		(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)
corp_corr							-0.010
							(0.096)
index_removal			-0.041***	-0.040***	-0.041***	-0.041***	-0.041***
			(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
iscorruption:index_removal				-0.053***			
				(0.020)			
iscompany:index_removal					-0.004		
					(0.035)		
iscorr_title:index_removal						-0.070**	
						(0.032)	
corp_corr:index_removal							-0.100
							(0.239)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	YES	Yes	Yes	Yes	Yes	Yes	Yes
Observations	665,507	665,507	665,507	665,507	665,507	665,507	665,507
Adjusted R2	0.119	0.119	0.119	0.119	0.119	0.119	0.119
Residual Std. Error	0.439	0.439	0.439	0.439	0.439	0.439	0.439
Note:							*p<0.1; **p<0.05; ***p<0.01

7 Discussion

Managerial und Policy Implications

In this chapter, I will relate the finding of the quantitative analysis to the relevant literature. The results are intended to build on existing literature and complement it with new insights into the phenomenon of citizen journalism. In contrast to the analysis, I start with the discussion on H2 and H3 to clarify what defines and differs citizen journalists from their "professional colleagues".

Citizen journalists are non-professional and therefore one can suggest that they publish on a more intrinsic purpose than their professional peers. With their journalistic work, they integrate into an existing system, however their interest and journalistic focus might differ compared to professional journalists. Furthermore, citizen journalists tend to draw on personal experience or events as a source for their reporting (Fogarty, 2012; Habermas, 1999; McCombs, 2004; Zelizer, 1993). Citizen journalists other than their professional colleagues mostly publish and engage on digital platforms which provide more room and leeway than printed journals whose content is limited and strongly determined by the editorial staff. In consequence of the advent of digital technologies, the number of potential readers exploded while competition in the field of citizen journalism has intensified in the same time (Goldfarb & Tucker, 2019; Shapiro & Varian, 1999; Waldfogel, 2017). To attract the high traffic of potential readership, a platform must offer broad topic coverage so that an online network can be consolidated. Anderson (2008) describes the random "long-tail" as a potential strategy to expand the network by an increasing cumulative probability of an eventual high popularity publishment, as articles can be successful by random (Peukert, 2019).

This research supports the statement with the confirmation of H1, concluding that the implementation of community-based article posts leads to a coverage of different topics than reporting conducted only by professional journalists. Online media "democratizes" the news and the public sphere. The active participation of citizens in news publishing can therefore reshape the news agenda (Brown & Waltzer, 2004). Even if citizen journalists receive less attention on average, in some cases citizen journalist show extreme high outliners in attracting attention, which highlights the relevance and success of their inclusion. The competition for readership attention constantly increases and lead to publications whose content is less individual but more convenient and quick to publish (Franck, 2019; Shapiro & Varian, 1999),

as investigative and qualitative news coverage is resource and capability intensive (Zelizer, 1993). Following the theory, I expect differences in the news coverage between author types due to individual influences and motivations. This research illustrates that citizen journalists do tend to cover more corporate topics. However, the characteristics of an article published by a citizen journalist differs in comparison to professional journalists. The title of an article is an important and first touchpoint with the readership, which is strongly attracted by media signals as this work can support with its research (Andrews & Caren, 2010). A monetary and in consequence attention driven professional journalism industry applies the use of corruption keywords more effectively and subsequently attracts more readership attention. In opposite to monetary incentives, the fact that the point index was removed, does noticeably influence the coefficients, which might be based on a perceived change in reward for publications. Citizen journalists do cover significantly more corporate news compared to professional journalists. Readership attention significantly increases when citizen journalists cover corporate topics, as readers do account higher for articles covering corporate topics from citizen journalists. The citizen authors might symbolize a person they can identify with and are therefore embodied as a reliable source (Reimers & Waldfogel, 2020). The demand side however accounts less for corruption articles published by citizen journalists. Corruption topics might require more investigation, expertise, resources and research (Zelizer, 1993).

Hence, the analysis of my sample confirms the hypothesis H2, that differences exist between citizen journalists and professional journalists.

I believe that making previously exclusive newspapers in which only professional journalists could publish accessible to citizen journalist is a reasonable economic consequence and reasonable managerial decision. This does not only mean including a comment function as it is established in most online newspapers, it rather entails the option for substantial publications of articles that are above the 280 character Twitter benchmark for citizens. The participatory public can enrich the offer for the reader and at the same time may ensure a stronger market presence through more attention for the online newspaper. This in turn translates into increased revenue for shareholders through monetized clicks. The risk of low quality publications under the publishing houses brand however needs to be critical observed by the management and if necessary controlled with minimum standards function as gatekeepers for postings.

Besides a broader network, the parameters that are attracting readership attention can be measured and exactly defined. Online media can extract massive amounts of data, conventional print media would not be able to measure (Peukert & Reimers, 2018).

As stated in my analysis, a non-monetary incentive system increases the quantity of items and affects the category. However, an unweighted point system rewards only quantity. I recommend an average score similar to common online store product rating methods to intrinsically incentivize unpaid participation on the platform.

Data Limitations

The panel data set gives this work the opportunity to perform extensive sample examinations in many different facets. However, I point out limitations which are underlying my investigation.

While one can generally explore attention through views in conjunction with my input variables, I do not have any other impression-inducing parameters such as geographic and demographic data for the readers. For a more granular study these variables would be interesting to investigate the readership in a more granular fashion. The article categories show significant differences in attention, but one cannot control for the place where the article is displayed on the website. Further, I may take into account that for example landing pages are by nature more intrinsically readable. My economic model does furthermore not take standard choices referring to Caplin (2016) into account. The reader is addressed by simultaneously, when choosing the article. Only views are counted, however, not in relation to the offer available in parallel. The variable *corp_corr* needs to be analyzed carefully as the sample I generate consists of 25 observations which may potentially underestimate the standard deviation, however research seems still in discussion for consents in terms of sample size (Kar & Ramalingam, 2013). I operationalize the text with search strings including companies mostly listed at the Austrian stock market.

In my theoretical analysis I find that, among other factors, an image of the author in his profile leads to more accountability. Although this possibility is available for authors on the website, I cannot control for due to missing variables in the data set.

Methodic limits mainly addressing the LPM model are previously described and taken care of. Limitations of this research have to take into account a large reduction in variation in my model due to the use of multiple fixed effects (Mummolo & Peterson, 2018). This might lead to a loss of important and significant insights further leading to a Type II error and consequently to mistakenly confirmation (Hill et al., 2020). In addition a probit regression to cross-check results is applied in table 5 appendix.

8 Conclusion

The above mentioned and discussed results of my work, demonstrate that citizen journalists have emerged to much more than just “hobby bloggers”. In their reporting, citizen journalists cover a wide range of topics, while their articles reach a considerable readership and in some areas even receive more attention than those of their professional colleagues. They play a pivotal role in addressing and presenting topics to the public by highlighting concerns and casting of additional topics.

Against this background, it is no longer possible to think of the media landscape without the contribution of citizen journalists. Their integration into traditional media houses should not only include already established formats such as reader letters but could also open up new opportunities and possibilities for publishing houses. Media houses can use platforms for citizen journalists to consolidate their offer in niche topics and thus their reach in a highly competitive market. Here, however, the aspect of journalistic quality control must always and absolutely be considered. In my analysis, I have not been able to trace the general worries that article quality might suffer as a result of the battle for readership attention in the past. However, I recognize in my research not only a significant impact of the inclusion of keywords on demand but also a call for more extended reports.

Large media houses play a central role in the formation of political and social debate as their "brands" have grown and represent a reliable quality journalism. If these brands provide a platform for "amateur" journalists, journalistic quality standards must be applied. Finding the right balance between being a journalistic gatekeeper without having too much influence on the topics and the positions of the actors will be one of the central challenges of the future.

In addition, future research might also focus on the regulatory impact of citizen journalist on the valuation of the company in question building on the work of Enikolopov et al. (2018) who examined the effects on the share price development of companies that are in the focus of citizen journalistic investigations.

9 Reference List

Agee, W. K., Ault, P. H., & Emery, E. (1983). *Reporting and writing the news*. Harper & Row.

Anderson, C. (2008). *The long tail: Why the future of business is selling less of more* (Rev. and updated ed). Hyperion.

Andrews, K. T., & Caren, N. (2010). Making the News: Movement Organizations, Media Attention, and the Public Agenda. *American Sociological Review*, *75*(6), 841–866.

Angrist, J., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. In *Mostly Harmless Econometrics: An Empiricist's Companion*.

Atton, C. (2015). Citizen Journalism. In W. Donsbach (Ed.), *The International Encyclopedia of Communication* (pp. 1–5). John Wiley & Sons, Ltd.

Barnhurst, K. G. (2013). Newspapers experiment online: Story content after a decade on the web. *Journalism: Theory, Practice & Criticism*, *14*(1), 3–21.

Baye, M. R., & Prince, J. (2014). *Managerial economics and business strategy* (Eighth edition). McGraw-Hill Irwin.

Berkowitz, D., & TerKeurst, J. V. (1999). Community as Interpretive Community: Rethinking the Journalist-Source Relationship. *Journal of Communication*, *49*(3), 125–136.

Bolter, J. D., & Grusin, R. A. (1999). *Remediation: Understanding new media*. MIT Press.

Bowman, S., & Willis, C. (2003). We media. *How Audiences Are Shaping the Future of News and Information*, *66*.

Brown, C., & Waltzer, H. (2004). Virtual sources: Organized interests and democratization by the Web. *The Social Science Journal*, *41*(4), 543–558.

Brunetti, A., & Weder, B. (2003). A free press is bad news for corruption. *Journal of Public Economics*, *87*(7–8), 1801–1824.

Brynjolfsson, E., Hu, Y. (Jeffrey), & Smith, M. D. (2003). Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. *Management Science*, *49*(11), 1580–1596.

Byström, K. (2002). Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology*, *53*(7), 581–591.

Cammaerts, B. (2011). Disruptive sharing in a digital age: Rejecting neoliberalism? *Continuum*, *25*(1), 47–62.

Campante, F. R., & Do, Q.-A. (2014). Isolated Capital Cities, Accountability, and Corruption: Evidence from US States. *American Economic Review*, *104*(8), 2456–2481.

Caplin, A. (2016). Measuring and Modeling Attention. *Annual Review of Economics*, *8*(1), 379–403.

Carr, D. J., Barnidge, M., Lee, B. G., & Tsang, S. J. (2014). Cynics and Skeptics: Evaluating the Credibility of Mainstream and Citizen Journalism. *Journalism & Mass Communication Quarterly*, *91*(3), 452–470.

Chiou, L., & Tucker, C. (2013). Paywalls and the demand for news. *Information Economics and Policy*, *25*(2), 61–69.

Chiou, L., & Tucker, C. (2017). Content aggregation by platforms: The case of the news media:. *Journal of Economics & Management Strategy*.

Claussen, J., Essling, C., & Peukert, C. (2018). Demand variation, strategic flexibility and market entry: Evidence from the U.S. airline industry. *Strategic Management Journal*, *39*(11), 2877–2898.

Coto-Millán, P. (2003). *Utility and Production*. Physica-Verlag HD.

Edwards, G. C., & Wood, B. D. (1999). Who Influences Whom? The President, Congress, and the Media. *American Political Science Review*, *93*(2), 327–344.

Enikolopov, R., Petrova, M., & Sonin, K. (2018). Social Media and Corruption. *American Economic Journal: Applied Economics*, *10*(1), 150–174.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in *R*. *Journal of Statistical Software*, *25*(5).

Fogarty, B. J. (2012). Local newspapers, House members, and source usage. *The Social Science Journal*, *49*(3), 275–283.

Franck, G. (2019). The economy of attention. *Journal of Sociology*, *55*(1), 8–19.

Franzosi, R. (1987). The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *20*(1), 5–16.

Gamson, W. A., Croteau, D., Hoynes, W., & Sasson, T. (1992). Media Images and the Social Construction of Reality. *Annual Review of Sociology*, *18*(1), 373–393.

Gardiner, J. (2017). Defining Corruption. In M. Johnston (Ed.), *Political Corruption* (3rd ed., pp. 25–40). Routledge.

Gentzkow, M., & Shapiro, J. M. (2006). Media Bias and Reputation. *Journal of Political Economy*, *114*(2), 280–316.

George, L. M., & Hogendorn, C. (2013). Local News Online: Aggregators, Geo-Targeting and the Market for Local News. *SSRN Electronic Journal*.

Gillmor, D. (2004). *We the media: Grassroots journalism by the people, for the people* (1st ed). O'Reilly.

Goldfarb, A., & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, *57*(1), 3–43.

Goode, L. (2009). Social news, citizen journalism and democracy. *New Media & Society*, *11*(8), 1287–1305.

Greene, W. (2004). Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model. *Econometric Reviews*, *23*(2), 125–147.

Habermas, J. (1999). *The Structural transformation of the public sphere: An inquiry into a category of bourgeois society* (10. print). MIT Press.

Healy, P. M., & Ramanna, K. (2013). When the Crowd Fights Corruption. *Harvard Business Review 91 (1/2): 122–29*.

Hill, T. D., Davis, A. P., Roos, J. M., & French, M. T. (2020). Limitations of Fixed-Effects Models for Panel Data. *Sociological Perspectives*, *63*(3), 357–369.

Hovland, C. I., & Weiss, W. (1951). The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, *15*(4), 635.

Johnson, K. A., & Wiedenbeck, S. (2009). Enhancing Perceived Credibility of Citizen Journalism Web Sites. *Journalism & Mass Communication Quarterly*, *86*(2), 332–348.

Johnson, R. A., & Sharma, S. (2004). The Struggle Against Corruption. In R. A. Johnson (Ed.), *The Struggle Against Corruption* (pp. 1–19). Palgrave Macmillan US.

Kar, S., & Ramalingam, A. (2013). Is 30 the magic number? Issues in sample size estimation. *National Journal of Community Medicine*, *4*, 175–179.

Kaufhold, K., Valenzuela, S., & de Zúñiga, H. G. (2010). Citizen Journalism and Democracy: How User-Generated News Use Relates to Political Knowledge and Participation. *Journalism & Mass Communication Quarterly*, *87*(3–4), 515–529.

Koopmans, R. (2004). Movements and media: Selection processes and evolutionary dynamics in the public sphere. *Theory and Society*, *33*(3/4), 367–391.

Kovach, B., & Rosenstiel, T. (2014). *The elements of journalism: What newspeople should know and the public should expect* (Revised and updated third edition). Three Rivers Press.

Krieghbaum, Hillier. (1956). *Facts in perspective ; the editorial page and news interpretation.* Prentice-Hall.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review*, *104*(8), 2421–2455.

McCarthy, J. D., McPhail, C., & Smith, J. (1996). Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991. *American Sociological Review*, *61*(3), 478.

McCombs, M. E. (2004). *Setting the agenda: The mass media and public opinion*. Polity ; Blackwell Pub.

McManus, J. H. (1994). *Market-driven journalism: Let the citizen beware?* Sage Publications.

Meinbezirk.at: Impressum.
Retrieved September, 2020, from: https://www.meinbezirk.at/s/impressum-oesterreich

Mohr, M. (2020). Politische Partizipation in Österreich 2019. Statista

Mummolo, J., & Peterson, E. (2018). Improving the Interpretation of Fixed Effects Regression Results. *Political Science Research and Methods*, *6*(4), 829–835.

OEWA Plus: Ergebnisse der Studie 2019-IV.
Retrieved September, 2020, from: http://www.oewa.at

Patterson, T. E. (1996). Bad News, Bad Governance. *The ANNALS of the American Academy of Political and Social Science*, *546*(1), 97–108.

Pennington, R. (2012). Mass media content as cultural theory. *The Social Science Journal*, *49*(1), 98–107.

Peukert, C. (2019). The next wave of digital technological change and the cultural industries. *Journal of Cultural Economics*, *43*(2), 189–210.

Peukert, C., & Reimers, I. (2018). Digital Disintermediation and Efficiency in the Market for Ideas. *SSRN Electronic Journal*.

Prat, A., & Strömberg, D. (2013). The Political Economy of Mass Media. In D. Acemoglu, M. Arellano, & E. Dekel (Eds.), *Advances in Economics and Econometrics* (pp. 135–187). Cambridge University Press.

Rayner, K., Miller, B., & Rotello, C. M. (2008). Eye movements when looking at print advertisements: The goal of the viewer matters. *Applied Cognitive Psychology*, *22*(5), 697–707. https://doi.org/10.1002/acp.1389

Reimers, I., & Waldfogel, J. (2020). *Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings* (No. w26776; p. w26776). National Bureau of Economic Research.

Regionalmedien Austria: Über uns.
Retrieved September, 2020, from: https://www.regionalmedien.at/ueber-uns/

Riano, J., Hodess, R., Heinrich, F., & Transparency International (Eds.). (2010). *Global corruption barometer 2010*.

Robinson, S., & DeShano, C. (2011). ‘Anyone can know’: Citizen journalism and the interpretive community of the mainstream press. *Journalism: Theory, Practice & Criticism*, *12*(8), 963–982.

Schudson, M. (1996). *The power of news*. Harvard University Press.

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Harvard Business School Press.

Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. Routledge.

Snyder, J. M., & Strömberg, D. (2010). Press Coverage and Political Accountability. *Journal of Political Economy*, *118*(2), 355–408.

Stark, A. (1997). Beyond Quid Pro Quo: What’s Wrong with Private Gain from Public Office? *American Political Science Review*, *91*(1), 108–120.

Transparency International: CORRUPTION PERCEPTIONS INDEX 2019 Retrieved August, 2020, from: https://www.transparency.org/en/cpi/2019/results/aut#details

Waldfogel, J. (2017). How Digitization Has Created a Golden Age of Music, Movies, Books, and Television. *Journal of Economic Perspectives*, *31*(3), 195–214.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (Sixth edition). Cengage Learning.

Zelizer, B. (1993). Journalists as interpretive communities. *Critical Studies in Mass Communication*, *10*(3), 219–237.

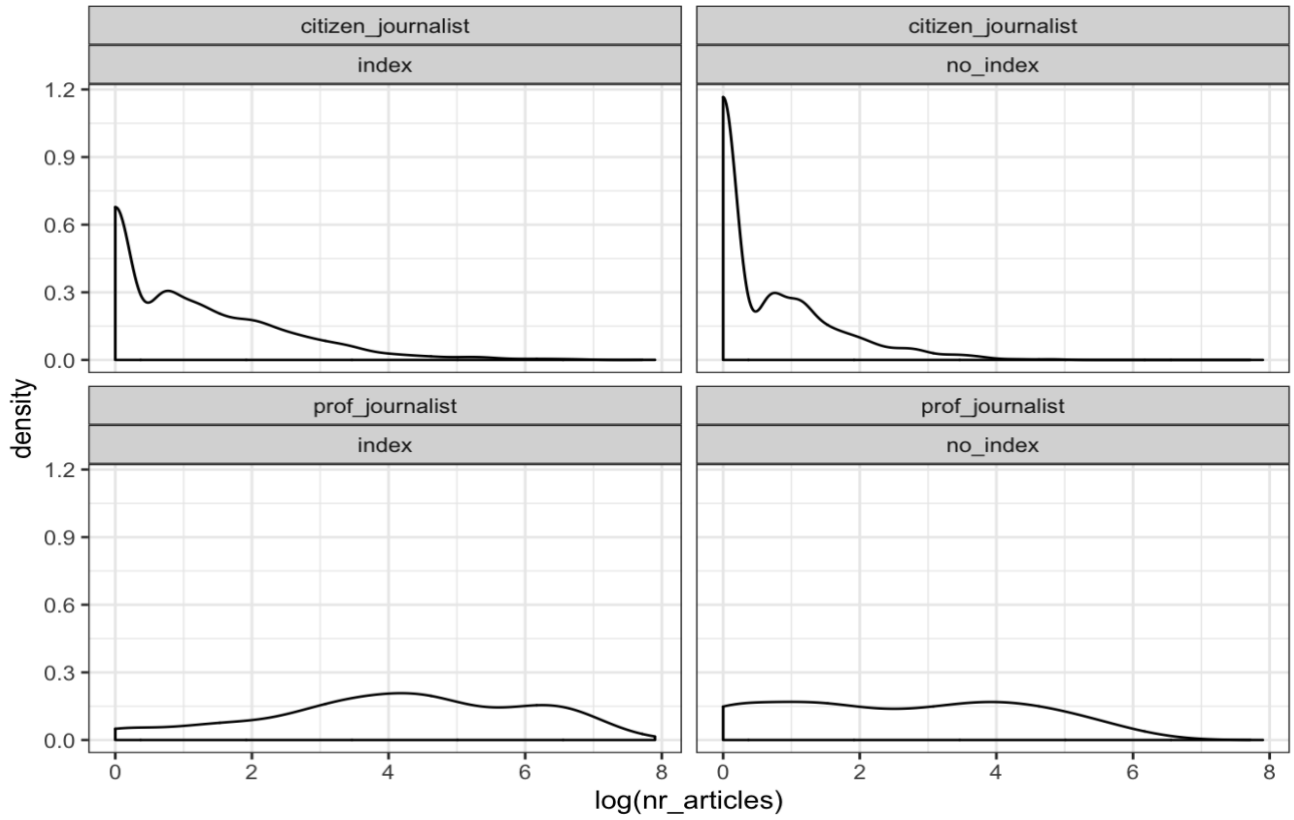
10 Appendix

10.1 Figures

Figure 3: Summary Statistics for Citizen Journalists (=1) & Index Removal (=1)

	0		1	
	0 (N=407063)	**1 (N=44087)**	**0 (N=203892)**	**1 (N=10465)**
#Views				
Mean (SD)	328 (1090)	262 (1160)	215 (824)	134 (560)
Median [Min, Max]	116 [0, 134000]	64.0 [1.00, 54700]	103 [0, 131000]	45.0 [0, 26800]
Words in Text				
Mean (SD)	179 (148)	192 (147)	167 (173)	162 (162)
Median [Min, Max]	141 [1.00, 4830]	152 [1.00, 2750]	124 [1.00, 8260]	119 [1.00, 2630]
no_pictures				
Mean (SD)	8.03 (26.9)	6.56 (22.4)	8.40 (23.7)	8.06 (20.2)
Median [Min, Max]	1.00 [0, 1310]	1.00 [0, 565]	2.00 [0, 924]	2.00 [0, 362]
no_comments				
Mean (SD)	0.0521 (0.465)	0.0614 (0.795)	0.578 (2.75)	0.960 (4.74)
Median [Min, Max]	0 [0, 44.0]	0 [0, 129]	0 [0, 84.0]	0 [0, 67.0]

Figure 4: Pre and Post Index Removal author productivity (no adaption time)



10.2 Tables

Table 3: (H1) Results for readership attention

Readership Attention log(Views)

	Dependent variable:					
	ln_no_views					
	(1)	(2)	(3)	(4)	(5)	(6)
log(words_in_text)	0.262***		0.261***	0.252***	0.250***	0.245***
	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)
ln_no_pictures	0.324***		0.325***	0.333***	0.353***	0.346***
	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)
log(words_in_title)	0.351***		0.352***	0.259***	0.268***	0.255***
	(0.003)		(0.003)	(0.003)	(0.003)	(0.003)
ln_no_comments	0.311***		0.311***	0.322***	0.289***	0.311***
	(0.004)		(0.004)	(0.004)	(0.004)	(0.004)
iscorruption		0.268***	0.172***	0.173***	0.159***	0.149***
		(0.015)	(0.014)	(0.013)	(0.013)	(0.013)
iscompany		0.111***	-0.064**	0.006	-0.033	0.011
		(0.028)	(0.025)	(0.025)	(0.025)	(0.024)
iscorr_title		0.176***	0.557***	0.591***	0.563***	0.565***
		(0.029)	(0.027)	(0.026)	(0.026)	(0.026)
iscorruption:iscompany		-0.375			-0.357	-0.380*
		(0.245)			(0.217)	(0.216)
Article Category FE	NO	NO	NO	NO	Yes	Yes
State FE	NO	NO	NO	Yes	NO	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	665,507	665,507	665,507	665,507	665,507	665,507
Adjusted R2	0.189	0.041	0.189	0.240	0.246	0.258
Residual Std. Error	1.118	1.215	1.118	1.082	1.078	1.069

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

Hypothesis 1

Table 4: (H4) Results for readership attention

Readership Attention log(Views)

	Dependent variable:						
	ln_no_views						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
citizen_journalist	-0.283***	-0.225***	-0.283***	-0.234***	-0.262***	-0.264***	-0.263***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
iscorruption			0.259***	0.245***	0.173***	0.147***	
			(0.015)	(0.014)	(0.016)	(0.013)	
iscompany			0.099***	0.170***	0.013	-0.024	
			(0.027)	(0.026)	(0.024)	(0.029)	
corp_corr							-0.110
							(0.244)
iscorr_title			0.154***	0.209***	0.545***	0.547***	0.600***
			(0.029)	(0.028)	(0.026)	(0.026)	(0.025)
log(words_in_text)					0.236***	0.236***	0.237***
					(0.002)	(0.002)	(0.002)
ln_no_pictures					0.352***	0.352***	0.352***
					(0.001)	(0.001)	(0.001)
log(words_in_title)					0.246***	0.246***	0.246***
					(0.003)	(0.003)	(0.003)
ln_no_comments					0.363***	0.364***	0.364***
					(0.004)	(0.004)	(0.004)
citizen_journalist:iscorruption					-0.087***		
					(0.029)		
citizen_journalist:iscompany						0.127**	
						(0.053)	
citizen_journalist:corp_corr							-0.502
							(0.498)
Article Category FE	NO	Yes	NO	Yes	Yes	Yes	Yes
State FE	NO	NO	NO	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	665,507	665,507	665,507	665,507	665,507	665,507	665,507
Adjusted R2	0.052	0.112	0.052	0.132	0.267	0.267	0.267
Residual Std. Error	1.209	1.170	1.208	1.157	1.063	1.063	1.063

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

Hypothesis 4

Table 5: Probit Model with output “Citizen Journalist”

	Dependent variable:			
	citizen_journalist			
	(1)	(2)	(3)	(4)
factor(category)local	0.078***	0.078***	0.078***	0.078***
	(0.006)	(0.006)	(0.006)	(0.006)
factor(category)news	-4.495	-4.495	-4.495	-4.496
	(3.510)	(3.510)	(3.510)	(3.510)
factor(category)people	0.027***	0.027***	0.027***	0.027***
	(0.007)	(0.007)	(0.007)	(0.007)
factor(category)politics	0.149***	0.149***	0.149***	0.149***
	(0.009)	(0.009)	(0.009)	(0.009)
factor(category)sports	0.194***	0.194***	0.194***	0.193***
	(0.007)	(0.007)	(0.007)	(0.007)
log(words_in_text)	-0.130***	-0.130***	-0.130***	-0.130***
	(0.002)	(0.002)	(0.002)	(0.002)
ln_no_pictures	0.075***	0.075***	0.075***	0.075***
	(0.002)	(0.002)	(0.002)	(0.002)
log(words_in_title)	-0.144***	-0.143***	-0.144***	-0.144***
	(0.004)	(0.004)	(0.004)	(0.004)
ln_no_comments	0.747***	0.747***	0.747***	0.747***
	(0.006)	(0.006)	(0.006)	(0.006)
iscorruption	0.032*	0.017	0.017	
	(0.017)	(0.017)	(0.017)	
corp_corr				-0.033
				(0.298)
index_removal	-0.412***	-0.414***	-0.413***	-0.414***
	(0.007)	(0.007)	(0.007)	(0.007)
iscorr_title:index_removal			-0.538***	
			(0.148)	
iscompany	0.053*	0.055*	0.053*	
	(0.030)	(0.032)	(0.030)	
iscorr_title	-0.266***	-0.268***	-0.231***	-0.262***
	(0.034)	(0.034)	(0.035)	(0.033)
iscorruption:index_removal	-0.273***			
	(0.075)			
iscompany:index_removal		-0.036		
		(0.121)		
corp_corr:index_removal				-4.078
				(185.421)
Constant	0.217***	0.217***	0.217***	0.217***
	(0.011)	(0.011)	(0.011)	(0.011)
Observations	665,507	665,507	665,507	665,507
Log Likelihood	-393,466.300	-393,473.400	-393,465.700	-393,474.900
Akaike Inf. Crit.	786,962.700	786,976.700	786,961.400	786,977.800
Note:			*p<0.1; **p<0.05; ***p<0.01	
			Probit Model Summarized	

Table 6: Alternative Model for Corporate Misconduct (business_cat*iscorruption)

	Dependent variable:		
	ln_no_views		citizen_journalist
	(1)	(2)	(3)
citizen_journalist		-0.289***	
		(0.003)	
business_cat	0.051***	0.051***	-0.026***
	(0.005)	(0.005)	(0.002)
iscorruption	0.226***	0.226***	-0.003
	(0.014)	(0.014)	(0.006)
index_removal			-0.122***
			(0.002)
iscorr_title	0.561***	0.561***	-0.077***
	(0.026)	(0.026)	(0.011)
log(words_in_text)	0.241***	0.241***	-0.035***
	(0.002)	(0.002)	(0.001)
ln_no_pictures	0.341***	0.341***	0.023***
	(0.001)	(0.001)	(0.001)
log(words_in_title)	0.248***	0.248***	-0.042***
	(0.003)	(0.003)	(0.001)
ln_no_comments	0.381***	0.381***	0.204***
	(0.004)	(0.004)	(0.002)
business_cat:iscorruption	-0.457***	-0.457***	0.034*
	(0.042)	(0.042)	(0.017)
Observations	665,507	665,507	665,507
R2	0.251	0.251	0.101
Adjusted R2	0.251	0.251	0.101
Residual Std. Error	1.074 (df = 665484)	1.074 (df = 665484)	0.443 (df = 665488)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 7: Patterns of Corruption (R. A. Johnson & Sharma, 2004, p. 2)

1	bribery and graft (extortion and kickbacks
2	kleptocracy (stealing and privatizing public funds)
3	misappropriation (forgery, embezzlement, misuse of public funds)
4	non-performance of duties (cronyism)
5	influence-peddling (favor-brokering and conflict of interest);
6	acceptance of improper gifts (‘speed’ money);
7	protecting maladministration (cover-ups, perjury)
8	abuse of power (intimidation and torture)
9	manipulation of regulations (bias and election rigging);
10	electoral malpractice (vote buying and election rigging);
11	rent-seeking (public officials who illegally charge for services after creating an artificial shortage);
12	clientelism and patronage (politicians giving material favors in exchange for citizen support);
13	illegal campaign contributions (giving unregulated gifts to influence policies and regulations)

Table 8: Variable Explanation

Variable	Definition	Logarithm
author_id	Individual Id of author on platform	
late_adopt	Author joined platform after index was removed	
article_id	Individual Id of article on platform	
state	State of publishing	
region	Region of publishing	
time variables	time_published, publ_year, publ_month, publ_day, publ_hour	
category	Category under which article was published	
title	Content of article title in words	
abstract	Content of article abstract in words	
text	Content of article text in words	
no_pictures	Number of pictures included in article	x
title_length	Number of words in article text	
abstract_length	Number of words in article abstract	
author_town	Town of journalists origin	
author_type	Category of author profession	
citizen_journalist	If journalist is citizen journalist or not	
no_views	Number of article views	x
no_comments	Number of article comments	x
no_likes	Number of article likes	x
index_removal	The time, when index was removed	
Iscompany	Article texts contains company name	
Iscorruption	Article texts contains corruption keyword	
iscorr_title	Article title contains corruption keyword	
corp_corr	Article text contains corruption keyword & company keyword	

10.3 Reproducible Code

All codes for further review are attached separately to the work. The data are conducted in RStudio and the code is separated into four parts:
1. Creation of Data set
2. Text Mining
3. Regression Analysis
4. Descriptive Analysis

I have send all codes to my supervisor Christian Peukert, however you can reach out to me in terms of questions, ideas and suggestions: *jonaslueder@me.com*

10.4 Text Mining Code Snippets

```
#Cleaning Strings for special characters
```{r}
Clean_String <- function(string){
   # Lowercase
   temp <- tolower(string)
   # Remove everything that is not a number or letter .
   temp <- stringr::str_replace_all(temp,"[^a-zA-ZäöüßÄÖÜ\\s]", " ")
   # Shrink down to just one white space
   temp <- stringr::str_replace_all(temp,"[\\s]+", " ")
}

dt.thesis_companies$Emittent <- Clean_String(dt.thesis_companies$Emittent)
dt.thesis_keywords$Keywords <- Clean_String(dt.thesis_keywords$Keywords)
```
#Create variable and count for string (company keyword) in substring (text)
```{r}
df.keywordcheck <- data.frame(dt.thesis.mining)
for (name in dt.thesis_companies$Emittent){
     df.keywordcheck[, ncol(df.keywordcheck) + 1] <- str_count(dt.thesis.mining$text, name)
     names(df.keywordcheck)[ncol(df.keywordcheck)] <- paste0(name)
}
df.keywordcheck
```
#Corruption Keywords
```{r}
df.keywordcheck <- data.frame(dt.keywordcheck)
for (namekey in dt.thesis_keywords$Keywords){
     df.keywordcheck[, ncol(df.keywordcheck) + 1] <- str_count(dt.thesis.mining$text,
namekey)
     names(df.keywordcheck)[ncol(df.keywordcheck)] <- paste0(namekey)
}
df.keywordcheck
```
#Title corruption check
```{r include=FALSE}
df.keywordcheck <- data.frame(dt.keywordcheck)
for (namekey in dt.thesis_keywords$Keywords){
     df.keywordcheck[, ncol(df.keywordcheck) + 1] <- str_count(dt.thesis.mining$title,
namekey)
     names(df.keywordcheck)[ncol(df.keywordcheck)] <- paste0(namekey)
}
df.keywordcheck
```
```