"Intangible Capital Indicators Based on Web Scraping of Social Media"

Patrick Breithaupt

(ZEW Mannheim)

Reinhold Kesler

(University of Zurich)

Thomas Niebel

(ZEW Mannheim)

Christian Rammer

(ZEW Mannheim)

Paper prepared for the IARIW-ESCoE Conference
November 11-12, 2021
Session 1
Time: Thursday, November 11, 2021 [11:15-12:45 GMT+1]

Intangible Capital Indicators Based on Web Scraping of Social Media

Patrick Breithaupt[1] Reinhold Kesler[2] Thomas Niebel[1*] Christian Rammer[3]

September 2020

Abstract

Knowledge-based capital is a key factor for productivity growth. Over the past 15 years, it has been increasingly recognised that knowledge-based capital comprises much more than technological knowledge and that these other components are essential for understanding productivity developments and competitiveness of both firms and economies. We develop selected indicators for knowledge-based capital, often denoted as intangible capital, on the basis of publicly available data from online platforms. These indicators based on data from Facebook and the employer branding and review platform Kununu are compared by OLS regressions with firm-level survey data from the Mannheim Innovation Panel (MIP). All regressions show a positive and significant relationship between survey-based firm-level expenditures for marketing and on-the-job training and the respective information stemming from the online platforms. We therefore explore the possibility of predicting brand equity and firm-specific human capital with machine learning methods.

JEL-Classification: C81, E22, O30

Keywords: Web Scraping, Knowledge-Based Capital, Intangibles

[1] Digital Economy Research Department, ZEW – Leibniz Centre for European Economic Research

[2] Department of Business Administration - Chair for Entrepreneurship, University of Zurich

[3] Economics of Innovation and Industrial Dynamics Research Department, ZEW – Leibniz Centre for European Economic Research

[*] Corresponding author: *thomas.niebel@zew.de.*

Acknowledgments: The authors would like to thank the German Federal Ministry of Education and Research for providing funding for the research project (INFOWIK - Investments in New Forms of Knowledge-Based Capital; funding ID: 161FI007). Tobias Gließner provided excellent research assistance. All remaining errors are ours alone. For further information on the authors' other projects see www.zew.de/staff_pbr, www.rkesler.com, www.zew.de/staff_tni, www.zew.de/staff_cra, as well as the ZEW annual report on www.zew.de/en.

1 Introduction

Knowledge-based capital is a key factor for productivity growth. Over the past 15 years, it has been increasingly recognised that knowledge-based capital comprises much more than technological knowledge and that these other knowledge components are essential for understanding productivity developments and competitiveness of both firms and economic aggregates (sectors, regions, and economies). In the tradition of the new growth theory (Romer, 1986, 1990; Lucas, 1988) knowledge-based capital, also denoted as intangible capital, is often measured by the stock of technological knowledge and is approximated by accumulated R&D expenditure or the stock of patents.

Corrado et al. (2005, 2009) have proposed a classification of intangible capital goods that comprises three main components: (1) *innovative property*, (2) *computerised information*, and (3) *economic competencies*. While the first two components are already covered by different statistical surveys (R&D survey on technical knowledge, innovation survey on technical and non-technical innovation-related knowledge, investment surveys on expenditure on computerised information such as software and databases), comprehensive statistical data on *economic competencies* are scarce. These competencies include in particular *firm-specific human capital*, *organisational capital*, as well as *brand equity*.

In this paper, we describe a new way of measuring investments in *economic competencies* that do not require firm surveys but are calculated on the basis of publicly available data from online platforms. We focus on two types of economic competencies: investments in *brand equity* and investments in *firm-specific human capital*. For *brand equity*, we use the number of "likes" of a company on Facebook as our indicator. Individual ratings (by employees) on the employer branding and review platform Kununu provide information for both the "company image" (*brand equity*) and on-the-job training/career development (*firm-specific human capital*). Both platforms are market leaders in their respective segment in Germany. Compared to survey-based data, publicly available platform data provide a much broader coverage at substantially lower costs, a much higher timeliness, and a much higher frequency.

However, the quality of platform data might be contested. In order to provide a first test of data validity, we compare the two newly developed indicators with survey-based expenditures on marketing (*brand equity*) and on-the-job training (*firm-specific human capital*), using data from the Mannheim Innovation Panel (MIP), which is the German part of the Community Innovation Survey of the European Commission. The results show a positive and significant relationship between firm-level expenditures for marketing and on-the-job training and the respective information stemming from the online platforms Facebook and Kununu. We therefore explore the possibility of predicting *brand equity* and *firm-specific human capital* with machine learning methods. However, the (additional) explanatory power of the platform data is limited.

The rest of the paper is structured as follows: Section 2 provides an overview of the economic literature on intangible capital as well as on the literature on using platform-based data for economic research. Section 3 introduces our data collection approach and the survey data for comparison and provides descriptive statistics of our estimation sample. The empirical approach and the results of our OLS regressions comparing the platform and the survey data are presented in Sections 4 and 5. Section 6 explores the possibility of predicting firm-level intangible capital expenditures with machine learning methods. Section 7 concludes.

2 Literature Review

Our research relates to the ongoing efforts in improving the measurement of knowledge-based capital. The terms knowledge-based capital and intangible capital are used as synonyms in this strand of the literature. Research related to intangible capital was largely initiated by the seminal papers of Corrado et al. (2005, 2009), which proposed a framework for the classification of intangible capital.

On the sectoral and total economy level, a large number of studies has been released in the past ten years trying to improve the measurement of intangible capital and more importantly also analysing the economic impact of intangible capital. Notable contributions with respect to the economic implications of intangible capital are amongst others Corrado et al. (2013), Roth and Thum (2013), Chen et al. (2016), Niebel et al. (2017), Corrado et al. (2017), Chen (2018) and Adarov and Stehrer (2019). Roth (2019) offers a recent review of the literature, while Haskel and Westlake (2018) provide a more comprehensive overview of the topic.

Apart from measuring intangibles at the sectoral and total economy level, a number of firm-level surveys with special focus on intangibles were conducted (Awano et al., 2010a; Awano et al., 2010b; Perani and Guerrazzi, 2012; European Commission, 2014). Furthermore, there exists a number of studies analysing the impact of knowledge-based capital on firm performance based on pre-existing general firm surveys (Crass et al., 2014; Di Ubaldo and Siedschlag, 2020; Rammer et al., 2020).

The paper also relates to the growing literature of using web-scraped data for economic research. Claussen and Peukert (2019) show a strong increase in articles published in journals on the Financial Times 50 list between 2000 and 2018 that use data crawling for different use cases with data obtained from online platforms. Specifically, with the availability of many potential data sources on the Internet and a growing computing power, the possibility of viable web-based indicators has expanded in the last decades. For example, Ginsberg et al. (2009) use Google search query data to predict influenza-like disease activity in the United States. Similarly, Choi and Varian (2012) use search engine data to develop a set of economic indicators, e.g., for unemployment claims. Besides search query data being a viable predictor for a wide range of outcomes nowadays (see Gentzkow et al. (2019) for a brief overview on

nowcasting), other studies specifically leverage online platform data to approximate and predict economic outcomes. For instance, restaurant data from Yelp has been employed to measure local business activity, neighborhoods' socioeconomic characteristics, and consumption patterns (Glaeser et al., 2018; Dong et al., 2019; Davis et al., 2019).

In recent years there has also been a lot of research in the field of web-based innovation indicators. For example, Gök et al. (2015) develop an indicator for R&D activities based on website data. Similarly, Kinne and Lenz (2019) as well as Pukelis and Stanciauskas (2019) use texts on firm websites to create a statistical model to predict a company's innovation status. Axenbeck and Breithaupt (2019) investigate the relationship between a wide variety of firm website characteristics and the firm innovation status. In addition, Krüger et al. (2020) make use of texts and hyperlinks on firm websites, create an inter-firm network and investigate its relationship with firm innovativeness.

Social media data have also been used to analyse brand equity activities of firms, as social media have become a key channel for marketing and customer interaction (Bruhn et al., 2012). While many studies aim at deriving insights on firms' marketing performance (see Misirlis and Vlachopoulou, 2018) or to assess the use of social media by firms (see Arora et al., 2014), fewer studies are linked to the subject of this paper, to derive a measure of brand equity at the firm level. For example, Coursaris et al. (2016) calculate an engagement score based on the number of likes, comments and shares on companies' posts on Facebook and find that the engagement level has a positive effect on brand equity. Chung et al. (2015) use information (posts, comments, likes) from Facebook pages of 100 large Korean firms and demonstrate that these indicators are positively related to market performance. Tirunillai and Tellis (2012) use indicators on chatter activities (product reviews on websites) for fifteen firms and find that the volume of chatter has a strong positive relationship with firms' returns. Luo et al. (2013) manually classify web blogs on nine large firms from the computer and software industry in terms of positive or negative sentiment along with blog volume data and find a strong leading effect of this brand indicator on firm equity value. All these studies focus on a relatively small number of large firms since large firms tend to be much more engaged on social media than small and medium-sized firms. In our study, we add to the literature by deriving social media based indicators for a large number of firms across all industries and size classes.

Using data from professional networking platforms such as LinkedIn or employer review platforms (such as Glassdoor, RateMyEmployer or JobAdviser) to assess firms' human capital is much less frequent. Most works using this type of social media data focus on its role for employees (see Aguado et al., 2019), employee response to firm events (see Gortmaker et al., 2020) and recruiting (see Chiang and Suen, 2015; Zide et al., 2014) rather than a measure of employers' human capital. Ji et al. (2017) use data from the employer review platform Glassdoor to derive indicators on job satisfaction and find a positive association with lower financial reporting risk. Pisano et al. (2017) use LinkedIn data to analyse whether ownership concentration affects the disclosure of human capital information via social media

platforms. Banerji and Reimer (2019) analyse the social connections of firm founders based on LinkedIn information to investigate the impact of connectedness (as a specific indicator of firm-specific human capital) on funds raised and find a strong positive relationship.

In this paper, we contribute to the literature in two ways. First, we describe a fairly generalizable method for matching and linking firm-level survey data and platform-based data. Second, using publicly available information from social media, we are able to derive new indicators of firm-specific human capital and brand equity that can complement firm surveys, thus improving the measurement of knowledge-based capital.

3 Data and Descriptive Statistics

3.1 Data Collection

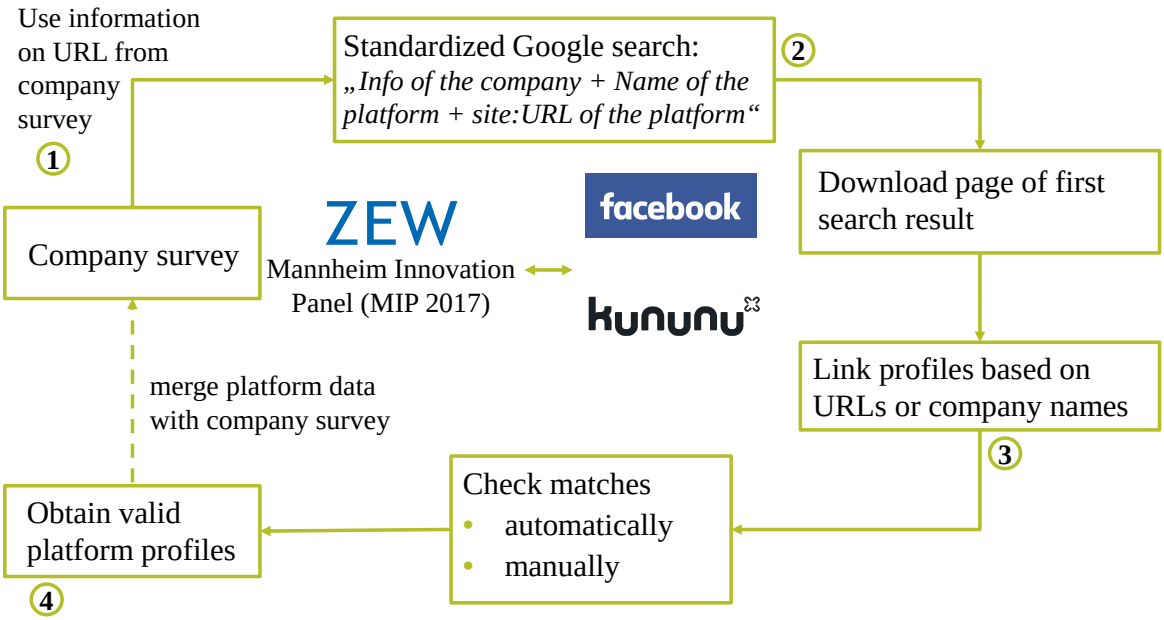
3.1.1 Identifying Platform Profiles

The aim of our data collection effort is to identify company information related to *brand equity* and *firm-specific human capital* from digital platforms (Facebook and Kununu[1]) and compare this information with survey-based indicators on the two aspects of firms' economic competencies taken from the German Innovation Survey (the "Mannheim Innovation Panel" - MIP). The MIP is the German part of the Community Innovation Survey (CIS) coordinated by Eurostat. The survey rests on a stratified random sample and is representative for the entire population of German enterprises (see Peters and Rammer, 2013).

Our data collection strategy is illustrated in Figure 3-1. For all firms that participated in the MIP survey conducted in 2017 (1) we create a specific search in Google based on the firm's website URL (2) to derive the platform URL of each company for the two platforms, Facebook and Kununu. We download the page behind the URL and check whether it is really the company from the survey. If there is a match (3), we can analyse the platform profiles (4).

[1] We also scraped the MIP 2017 company profiles on Twitter. The number of firms on Twitter is rather low, so we decided to not use Twitter data.

Figure 3-1: Data Collection Approach



The Google search included the company's website URL, the platform name and the search operator "*site:platform URL*", which only returns results from the platform page. This approach only necessitates some identifying information of a company and allows to be generally applied to different online platforms. We assume to be so specific that the first search result must be the platform profile of the searched company - unless there is none. Subsequently, we take the received platform URL of the first search result and download the HTML code of the page, which is often the start page of the company on the respective platform. Finally, we extract the initially selected company's key information, which should ideally also be on the platform page, link it to the survey and verify the match.

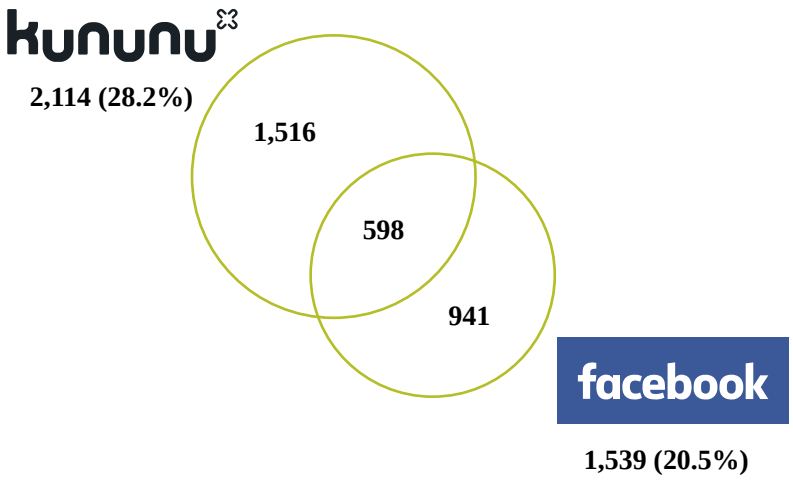
The sample of our analysis includes 7,498 companies.[2] To verify whether our search described in Figure 3-1 identified the right company on the platform, we compare information from the platform profile with the information from the survey, such as the company name or website URL. For Facebook, this is straightforward since companies are obliged to state their website URL on the start page. On the Kununu site of a company, which is not managed by the company, there is no corresponding imprint. We therefore analyse the similarity of the company name on Kununu and from the MIP. In a first step, we do exact string matching. If no exact match was established, the Python package *fuzzywuzzy* is used to perform a fuzzy string matching. For this purpose, we use the *fuzzywuzzy* functions *ratio*, *partial_ratio*, *token_sort_ratio* and *token_set_ratio* and equally weight the results. In addition, a minimum threshold

[2] In total, 8,278 firms participated in the MIP 2017. For about 9% of the firms, no URL of the company website was available. A first search result on Google produced Facebook URLs for 7,330 firms and Kununu URLs for 4,759 firms.

of 50 percent is defined. If multiple entries are above the threshold, we choose the MIP entry with the highest fuzzy matching score.

The search was carried out at the end of 2017. We obtained 2,114 company platform profiles for Kununu and 1,539 for Facebook, representing 28.2% and 20.5%, respectively, of our sample (see Figure 3-2). In the case of Facebook, the share is comparable to other studies based on the retrieval of corporate profiles on online platforms (see Bertschek and Kesler, 2018). For 598 firms, we found both a Kununu and a Facebook page.

Figure 3-2: Identified Platform Profiles

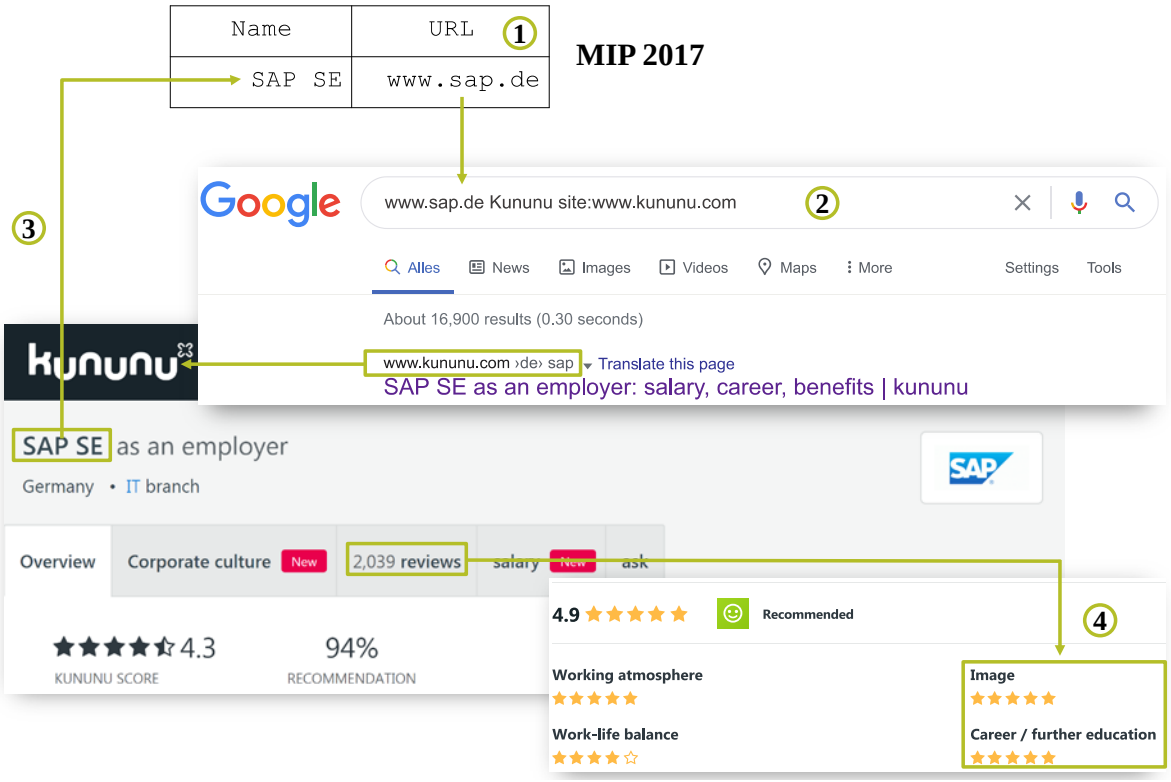


Note: In total, there are 7,498 companies with valid URLs for the company website that participated in the MIP 2017 survey. Out of these 7,498 companies, we were able to identify 28.2% companies with a Kununu page and 20.5% companies with a Facebook page.

3.1.2 Obtaining Kununu Ratings for Training and Image

Kununu is an employer branding and review platform founded in 2007 and was acquired by Xing (a German competitor of LinkedIn) in 2013. Albeit having a dedicated website for companies in the U.S., Kununu has a strong focus on German-speaking countries (Germany, Austria, and Switzerland). Besides an overall score/rating, employees can evaluate their company within different categories. For our purposes, the individual ratings for “company image” and “on-the-job training/career development” are the relevant categories (see Figure 3-3). “Company image” is within the framework for intangible capital by Corrado et al. (2005, 2009) related to *brand equity* as it reflects the firm's public image, which is a major factor for a firm's marketing success. The rating for “on-the-job training/career development” is directly linked to *firm-specific human capital* as it evaluates the relevance and effectiveness of a firm's human capital development efforts from the employees' point of view. Kununu data were collected in August 2018 with historical data back to 2010 (see Table A-1 in Appendix A).

Figure 3-3: Example of the Methodology Using the Kununu Platform

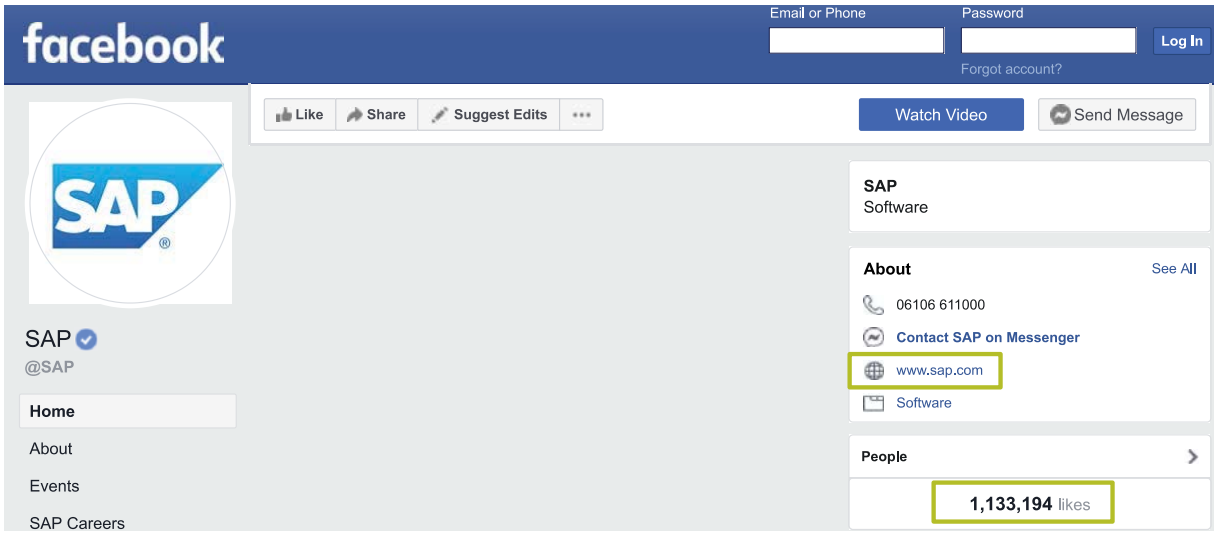


Note: This is just an example for a Kununu page of a firm. It does not necessarily mean that the firm is in our sample. Screenshots of the Kununu page are in German language and were automatically translated by Google Translate. The numbers 1-4 refer to the numbers in Figure 3-1.

3.1.3 Obtaining Facebook Likes

Data from Facebook were collected in December 2017 and the first week of January 2018. The relevant data on each of the companies' Facebook pages are the number of “likes” and the URL of the website of the company. The latter is needed to check whether our Google search indeed identified the right company and to merge the number of “likes” on the Facebook page with the survey data on intangibles and other company characteristics in the MIP 2017 survey. Within the framework of intangible capital by Corrado et al. (2005, 2009), the number of "likes" is related to *brand equity* as it reflects positive values associated with a firm in the general public. A high number of “likes” indicates that the firm's efforts to establish a favourable perception of its activities, products and services have been - at least to some extent - successful. We scraped the start page of the company profile on Facebook which includes the number of “likes” and the URL of the company website (see Figure 3-4). Obtaining historical Facebook data was not possible since Facebook has massively restricted API access (see Table A-1 in Appendix A) as a result of the Facebook–Cambridge Analytica data scandal in early 2018.

Figure 3-4: Example of the Company Profile on the Facebook Platform



Note: This is just an example for a Facebook page of a firm. It does not necessarily mean that this firm is also in our sample. Some elements of the Facebook profile of the company were manually removed for the sake of clarity.

3.2 Survey Data: Mannheim Innovation Panel (MIP)

We use data from the Mannheim Innovation Panel (MIP) to test the relevance of our measures of *brand equity* and *firm-specific human capital* derived from information provided on online platforms. The MIP is the German contribution to the Community Innovation Survey (CIS) of the European Commission and follows the survey methodology of the CIS. The MIP sample is a stratified random sample of about 13 percent of the target population, which includes firms with 5 or more employees from manufacturing and business-oriented services. The response rate of the 2017 survey was 25 percent, resulting in 8,278 observations. For a more detailed description of the survey see Peters and Rammer (2013) and Behrens et al. (2017).

For our analysis, two variables from the MIP 2017 survey are used, the amount of expenditure for marketing in 2016, and the amount of expenditure for employee training in 2016. Marketing expenditures include all in-house and contracted out expenditures for advertising and branding (incl. commercial marketing), reputation building, conceptual design of marketing strategies, market and costumer research, and the installation of new distribution channels. Pure selling costs are not considered as marketing expenditures. Employee training expenditures include all in-house and contracted out expenditures for training and further education of employees, including payroll costs of employees for working time used to attend training. Expenditures for vocational education are not part of training expenditures. 6,339 firms provided data on their marketing expenditures, and 6,419 reported the amount of training expenditures.

3.3 Descriptive Statistics

3.3.1 Kununu Data

Table 3-1 and Table 3-2 show the summary statistics for our estimation sample for the analysis of the relationship between our knowledge-based capital indicators stemming from the employer branding and review platform Kununu and the MIP 2017 survey-based expenditures for knowledge-based capital.

We restrict our sample to company profiles with at least four ratings between January 2017 and August 2018 as there is a trade-off between data quality and number of data points (i.e., more ratings per company on Kununu implies better data quality, but fewer data points). This reduces the number of observations with Kununu data on "training" to 813 in the full sample (see Table A-2 in Appendix A) and 519 in the estimation sample (see Table 3-1). The number of observations with Kununu data on "company image" is reduced to 805 in the full sample (see Table A-2 in Appendix A) and 492 in the estimation sample (see Table 3-2). These numbers are much lower than the total number of firms of 8,278 participating in the MIP 2017 survey (see Table A-4), reflecting the fact that only a smaller part of the entire firm population is represented on the Kununu platform. Overall, Industry J (Information and Communication) is over-represented in our estimation sample compared to the total MIP 2017 sample (see Table A-4). Furthermore, we do have much fewer firms with less than 10 employees in our sample compared to the full MIP 2017 sample (see Table A-5). This is driven by the fact that the ratings on Kununu are coming from the employees of the firm. For firms with less than 10 employees it is less likely to reach our minimum threshold of four ratings between January 2017 and August 2018.

Table 3-1: Summary Statistics – Training: Kununu Rating - Estimation Sample

	N	Mean	Median	SD	Min	Max
Training: Kununu rating	519	3.31	3.38	0.79	1	5
Training expenditures (MEUR)	519	1.10	0.060	13.8	0.00097	300
Turnover (MEUR)	519	404.2	27.5	2671.4	0.080	46800
Number of employees	519	1238.3	165	8614.6	1	122608
Number of Kununu ratings (Training)	519	22.9	9	74.3	4	1174

Note: These data contain only firms with at least four detailed ratings on Kununu between January 2017 and August 2018. "Number of Kununu ratings (Training)" is not part of the regressions in Section 5.1. Four is the lower threshold for the number of ratings. All firms with less than four ratings between January 2017 and August 2018 are not part of the empirical analyses.

Table 3-1 shows that the average rating for on-the-job training of the companies in our estimation sample is 3.31 (scale: 1-5; for more details see Figure B-1 and Figure B-4 in Appendix B), while the expenditures for on-the-job training stemming from the MIP 2017 survey are on average 1.1 million Euro. As there are very large firms with more than 122 thousand employees and more than 46 billion Euro in turnover in our sample, the average number of employees is at 1,238 and the average annual turnover is about 400 million Euro. However, the respective median values are much lower. The average number of ratings of a company on Kununu in our sample for on-the-job training is 22.9. Table 3-2 displays the

descriptive statistics for our estimation sample for our measure of *brand equity* on Kununu, which is the rating of the current and past employees for the "company image". Compared to the rating for on-the-job training, the average assessment of the employees for the image of their company is noticeably larger (3.62 vs 3.31). Figure B-2 and Figure B-5 provide more details about the distribution of the ratings. The average number of ratings for "company image" is slightly lower than for "training" (21.7 vs 22.9).

Table 3-2: Summary Statistics - Image: Kununu Rating - Estimation Sample

	N	Mean	Median	SD	Min	Max
Image: Kununu rating	492	3.62	3.75	0.80	1	5
Marketing expenditures (MEUR)	492	6.98	0.11	81.4	0.0010	1480
Turnover (MEUR)	492	312.7	27.4	1765.2	0.22	25763
Number of employees	492	1014.4	158	7189.5	3	122608
Number of Kununu ratings (Image)	492	21.7	8.50	72.7	4	1171

Note: These data contain only firms with at least four detailed ratings on Kununu between January 2017 and August 2018. "Number of Kununu ratings (Training)" is not part of the regressions in Section 5.1. Four is the lower threshold for the number of ratings. All firms with less than four ratings between January 2017 and August 2018 are not part of the empirical analyses.

3.3.2 Facebook Data

Table 3-3 displays the summary statistics for our estimation sample for the analysis of the relationship between the number of Facebook "likes" and the MIP 2017 survey-based marketing expenditures. In total, we have 944 companies with non-zero and non-missing data in our estimation sample. A company in our sample has on average 9,684 Facebook "likes" and 2.4 million Euro of marketing expenditures[3]. Overall, the firms in our estimation sample are on average larger than in the entire MIP 2017 sample. Especially the size class of firms with 0 to 9 employees is under-represented in our sample as these firms are less likely to have a Facebook page (see Table A-7). Therefore, the average turnover in our estimation sample is 58.2 million Euro and each company has on average close to 220 employees. However, these average numbers are generally driven by very large companies. All median values are by a large amount lower than the mean (see Table 3-3, column 4),

Table 3-3: Summary Statistics - Image: Facebook Likes - Estimation Sample

	N	Mean	Median	SD	Min	Max
Image: Facebook likes	944	9683.8	224	98292.4	1	1702502
Marketing expenditures (MEUR)	944	2.40	0.032	48.8	0.00048	1480
Turnover (MEUR)	944	58.2	5	434.9	0.027	11630
Number of employees	944	218.4	40.5	1090.8	1	25247

Note: These data contain only firms with at least one like on Facebook as of December 2017/January 2018 and non-zero marketing expenditures in 2016.

[3] Figure B-3 and Figure B-6 in Appendix B provide a histogram for the number of "likes" and a scatterplot for the relationship between the number of "likes" and the marketing expenditures.

4 Empirical Approach

We analyse the relationship between our newly developed platform-based indicators for *brand equity* and *firm-specific human capital* and the MIP 2017 survey-based measures on marketing expenditure and on training expenditure. These knowledge-based assets belong, within the framework of Corrado et al. (2005, 2009), to the group of *economic competencies* which are usually not measured within official statistics. For our cross-sectional data[4], we analyse this relationship with standard OLS regressions containing a set of firm-level control variables:

$$\ln Y_{image/likes,i} = \beta_{exp} \ln(expenditures\_marketing)_{2016,i} + \boldsymbol{X_i\gamma} + e_i \quad (1)$$

$\ln Y_{likes,i}$ denotes the number of "likes" on Facebook in December 2017/January 2018, $\ln Y_{image,i}$ is the average rating for the period from January 2017 to August 2018 for the item "image" on Kununu. The vector of control variables $\boldsymbol{X_i}$ includes turnover, the number of employees, and a set of 23 industry dummies. The number of Facebook "likes" and the average score for the item "company image/career development" on Kununu are our indicators for the intangible asset *brand equity*.

Furthermore, we study the explanatory power of our indicator for *firm-specific human capital* via the following OLS regression:

$$\ln Y_{training,i} = \beta_{exp} \ln(expenditures\_training)_{2016,i} + \boldsymbol{X_i\gamma} + e_i \quad (2)$$

$Y_{training,i}$ is the average rating for on-the-job training for the period from January 2017 to August 2018 for the employer branding and review platform Kununu. The vector of control variables $\boldsymbol{X_i}$ includes once more turnover, the number of employees, and a set of 23 industry dummies.

As an extension to our OLS regressions, which are just evaluating the relationship between the platform- and the survey-based data, we employ a machine learning (ML) approach for predicting firm-level expenditures for marketing and on-the-job training based on our platform indicators. Details can be found in Section 6.

[4] We gathered historical Kununu data for the years prior to 2017. Thus, in principle, it would be possible to do fixed effects panel regressions with the full MIP panel and the historical Kununu platform data. However, due to the limited number of observation/ratings in early years on Kununu, this was, after all, not feasible.

5 Estimation Results

5.1 Kununu

Table 5-1 shows the results for our OLS regressions for the relationship between the MIP 2017 survey-based measures for knowledge-based capital and our platform-based indicators. We estimate four different models described in Equation (1) and Equation (2). In column (1), we regress the Kununu rating for training on the log of the survey-based expenditures for training and a set of control variables. Column (3) displays the analogous results for the company image. In columns (2) and (4), we use log transformations of the dependent variables, as the Kununu ratings are slightly skewed (see Figure B-1 and Figure B-2). As mentioned before, we restrict our sample analysing the Kununu data to company profiles with at least four ratings between January 2017 and August 2018. More ratings per company on Kununu imply better data quality. But, on the other hand, the number of observations in our regressions is dramatically reduced. Given the data, the lower bound of at least four ratings per company is in our view sensible. We also provide robustness checks with varying thresholds.

Table 5-1: OLS Regressions Kununu

Dependent Variable:	(1) Training: Kununu rating	(2) ln(Training: Kununu rating)	(3) Image: Kununu rating	(4) ln(Image: Kununu rating)
ln(Training expenditures)	0.0680*	0.0237*		
	(1.81)	(1.88)		
ln(Marketing expenditures)			0.0842***	0.0272***
			(3.13)	(3.19)
ln(Turnover)	0.0121	0.0103	-0.0407	-0.00946
	(0.28)	(0.68)	(-0.83)	(-0.61)
ln(Number of employees)	-0.0590	-0.0238	-0.0459	-0.0160
	(-1.06)	(-1.27)	(-0.78)	(-0.83)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.139	0.139	0.128	0.132
Observations	519	519	492	492

Robust t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

These robustness checks varying the minimum number of ratings can be found in Table A-8 and Table A-9 in the Appendix. Requiring at least 3 or 5 ratings does not qualitatively change the results. With a minimum of 6 ratings, training gets insignificant due to the reduced number of observations. A minimum of 10 ratings, leads to a further drop in the number of observations resulting in insignificant results for both the image and training. For another robustness check, we removed the Kununu ratings of ex-employees, as especially in case they were fired, the resulting ratings could have a bias. However, the removal of the ex-employees, does not change results fundamentally.

All columns indicate a positive and significant relationship between the survey-based expenditure measures and our platform-based indicators. The main difference between columns (1) and (2) vs (3) and (4) are the higher significance levels of the latter.

5.2 Facebook

Table 5-2 presents the results for our OLS regressions for the relationship between the MIP 2017 survey-based marketing expenditures and the number of Facebook "likes" in December 2017/January 2018. In column (1), we regress the log of the number of Facebook likes on the log of the survey-based marketing expenditures of the firm. In column (2), we add turnover and the number of employees as explanatory variables. In column (3), we additionally include industry dummies. As before with the Kununu data, we observe a positive and highly significant relationship between the survey-based expenditures for marketing and our platform-based indicator (corresponding to the number of Facebook "likes").

Table 5-2: OLS Regressions Facebook

Dependent Variable:	(1) ln(Image: Facebook likes)	(2) ln(Image: Facebook likes)	(3) ln(Image: Facebook likes)
ln(Marketing expenditures)	0.522***	0.454***	0.455***
	(15.67)	(8.23)	(8.34)
ln(Turnover)		0.0589	0.106
		(0.72)	(1.11)
ln(Number of employees)		0.0550	0.0178
		(0.65)	(0.19)
Industry dummies	No	No	Yes
adj. R^2	0.261	0.265	0.364
Observations	944	944	944

Robust t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

6 Prediction of Expenditures on Knowledge-Based Capital based on Machine Learning

In this section, we discuss the use of machine learning methods to predict internal firm expenditures with the help of pre-existing firm and platform data. The basic idea is that internal expenditures could be estimated using (semi-) public data. The major advantage of this approach is that the expenditures could be updated more regularly as the platform data is often updated. The firm data (number of employees, industry, turnover and expenditures) is based on the Mannheim Innovation Panel (MIP).

Our analysis is based on the estimation samples from Section 3.3. As feature variables of the machine learning models, we choose the turnover of the firm, number of employees, industry and scraped data from Kununu and Facebook. The target variable is the amount of firm-specific expenditure on “on-the-job training/career development” or “company image" from the MIP survey. We remove some “outlier” corporations from our samples, because in the worst case, all “outlier” corporations are randomly assigned to the test dataset. This is a problem that predominantly occurs in smaller datasets. Based on each target variable, the entries above the 99% percentile are removed. Additionally, the target variable is log-transformed for the machine learning, but the predictions are transformed back with the exponential function before the evaluation metrics are calculated. This step is useful as the target variable distributions are skewed. A standardization of the feature variables “number of employees” and “turnover” is performed. As a result, the variables have a mean of zero and standard deviation of one in the training dataset. The transformation is performed to improve the performance of neural networks. The training data standardization parameters are also used to standardize the test data in order to avoid a data leakage. As we use the standardization parameters of the training data, the mean and standard deviation of the test data may not be zero and one. The categorical industry information is converted into dummy variables as most machine learning methods can only work with numerical data. Several machine learning regression models are trained for our analysis: *Neural networks (NN), random forests (RF), k-nearest-neighbour (KNN)* and *support vector machines (SVM).* An overview of the methods is given in Friedman et al. (2001). For each model we perform a 10 fold cross validation to find suitable model parameters and more robust models. The corresponding model parameters in the software packages are listed in italics and brackets. The neural network has three dense layers with the sizes 16, 8, and 1, uses the “ReLu” and “linear” (last layer) activation function and the “mean absolute error” objective function. The optimizers “Adam”, “Adadelta” and “SGD” *(optimizer)* are considered. The random forest is trained with 1,000 and 5,000 trees (*n_estimators*), the “mean absolute error” (*criterion*), a maximum depth of 2, 5, 10 and 20 (*max_depth*) as well as “None” and “auto” as maximum number of features (*max_features*). The k-nearest neighbors method considers the nearest 1, 2, 5, 10, 25 and 50 data points (*n_neighbors*) and the weighting schemes “uniform” and “distance” (*weights*). The support vector machine considers the kernel “linear”, “rbf”, “sigmoid” and “poly” (*kernel*) to map the data into a high-dimensional vector space and the regularization parameters 1, 2 and 5 (*C*). Non-specified parameters are set to the

default options. The parameter spaces are in theory expandable to retrieve improved results. For the training 67% of the data was used and 33% was reserved for model testing. In addition, we perform five random test train splits in order to be able to make a statement about the sensitivity of the splits. For model training and evaluation, we use the Python packages *Keras* (Chollet, 2015) and *Scikit-learn* (Pedregosa et al., 2011). *Keras* is used as the neural network implementation and the other machine learning models are based on *Scikit-learn*. We use the Mean Absolute Error (MAE) metric as evaluation measure, because it is commonly used and can be interpreted directly. For example, a MAE of 0.1 means that there is an average absolute difference of 100,000 Euros between the actual expenditure and our prediction.

Table 6-1: Predictive power for training expenditures based on Kununu data

Features				Result		
Number of employees	Turnover	Industry	Training Kununu	N	Best model	MAE
			x	513	RF	0.22 (0.04)
x	x			513	RF	0.15 (0.03)
x	x	x		513	RF	0.14 (0.03)
x	x	x	x	513	RF	0.15 (0.03)

The results are based on the mean values of five random train-test splits. The standard errors for MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train dataset as prediction has a MAE of 0.31 and 0.22.

Note: Target variable: Training expenditures. Firms with zero marketing expenditure, turnover or employees are dropped (estimation sample criterions). Six outliers are dropped based on the 99th percentile of the target variable. All numbers are rounded on two decimal places. The MAE reports the best "Mean Absolute Error" value for the specified set of models.

Table 6-2: Predictive power for marketing expenditures based on Kununu data

Features				Result		
Number of employees	Turnover	Industry	Image Kununu	N	Best model	MAE
			x	487	KNN	1.04 (0.09)
x	x			487	SVM	0.88 (0.09)
x	x	x		487	RF	0.87 (0.08)
x	x	x	x	487	RF	0.86 (0.10)

The results are based on the mean values of five random train-test splits. The standard errors for MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train dataset as prediction has a MAE of 1.45 and 1.05.

Note: Target variable: Marketing expenditures. Firms with zero marketing expenditure, turnover or employees are dropped (estimation sample criterions). Five outliers are dropped based on the 99th percentile of the target variable. All numbers are rounded on two decimal places. The MAE reports the best "Mean Absolute Error" value for the specified set of models.

Table 6-3: Predictive power for marketing expenditures based on Facebook data

Features				Result		
Number of employees	Turnover	Industry	Image Facebook	N	Best model	MAE
			x	934	KNN	0.26 (0.02)
x	x			934	NN	0.23 (0.02)
x	x	x		934	NN	0.22 (0.02)
x	x	x	x	934	RF	0.21 (0.01)

The results are based on the mean values of five random train-test splits. The standard errors for MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train dataset as prediction has a MAE of 0.42 and 0.27.

Note: Target variable: Marketing expenditures. Firms with zero marketing expenditure, turnover or employees are dropped (estimation sample criterions). Ten outliers are dropped based on the 99th percentile of the target variable. All numbers are rounded on two decimal places. The MAE reports the best "Mean Absolute Error" value for the specified set of models.

Table 6-1 shows the predictive performance for training expenditures based on Kununu and MIP data. Kununu data alone is as good as a baseline model with MAE of 0.22, but the MIP data explains more of the data with MAE of 0.14. Combining both feature sets does not improve the performance. In this case, the web-scraped data worsens our predictions slightly. Table 6-2 shows the predictive performance for internal marketing expenditures based on Kununu and MIP data. The model based on web-scraped data is as good as the baseline model. The MIP data, on the other hand, can again explain more of the data with MAE of 0.87. As expected, combining both feature sets does only slightly affect the performance. Lastly, Table 6-3 shows the results for the prediction of the internal marketing expenditure based on the Facebook and MIP data. A model based on the platform data has a MAE of 0.26, but is outperformed by a model based on the MIP data. Combining both feature sets results in a model with MAE of 0.21. This suggests that the web-scraped data has a positive effect on our predictions.

The results are robust, in the sense, that we can see the same pattern across different predictions. Platform data alone has a relatively small or no predictive power as the baseline models yield similar MAE values. MIP data explains a higher amount of the data and outperforms the platform data. Combining platform data (Facebook or Kununu) with MIP data has at most a slight effect or no effect on the results.

The main problem in our analysis is the low number of observations as we are limited to firms with data on firm-level training or image expenditure. Unfortunately, the coverage of MIP firms on Kununu and Facebook is relatively low as illustrated in Figure 3-2. However, we expect a better performance with an increasing number of observations. Machine learning models based on small data sets are, to some extent, sensitive to sample splitting. For example, all large corporations could fall into the test dataset leading to non-robust results. Our results are therefore based on the mean values of five random train-test splits. The reported standard errors are in some cases relatively high (up to 0.10). Additionally, the data is not representative. For example, small companies are underrepresented. This can lead to problems in the generalizability of the machine learning approach. The information about the "best model"

should not be interpreted too extensively as multiple models often perform only slightly different. Random forests are known to have a relatively good performance on tabular data. Therefore, it is not surprising that random forest are in the most cases the best model. The performance could be further improved with modified random forest models, e.g. boosted trees. Random forests, on the other hand, have the major disadvantage that the methodology is based on weighted averages. A random forest can therefore never predict firm-level expenditures that lie outside the training set. In summary, the MIP and platform data can be used to a limited extent to estimate the internal expenditures of companies.

7 Conclusions and Future Research

This paper aimed at developing new indicators for intangible capital of firms on the basis of publicly available data from online platforms. These basic indicators for *brand equity* and *firm-specific human capital,* which are part of the intangibles framework developed by Corrado et al. (2005, 2009), were taken from the social media platform Facebook and the employer branding and review platform Kununu. We compare these indicators by means of OLS regressions with firm-level survey data on marketing and training expenditure taken from the German part of the Community Innovation Survey. All regressions show a positive and significant relationship between the firm-level expenditures for marketing and on-the-job training and the respective information stemming from the online platforms. Various robustness checks confirm the validity of the results.

However, there are also caveats with our current approach. Due to the limited presence of smaller firms on online platforms, we are currently predominantly capturing medium-sized and larger firms. Furthermore, although we do find a positive and significant relationship between our platform-based indicators and the survey-based numbers in our OLS regressions, predicting expenditures based on an explorative machine learning approach shows that the platform data alone have little or no predictive power.

Using data from online platforms can nevertheless provide a useful source for establishing firm-level indicators on intangible assets in the field of economic competencies, which are difficult to measure through surveys or from balance sheet data. But in order to better utilise this data source, more research is required. First, we need a better understanding of the dynamic relationship between activities on online platforms related to a firm's knowledge-based capital, and the actual firm activities to build up and maintain such capital. Secondly, comparative analysis of different platform data are needed to better assess the value of the information that can be derived from various platforms. Finally, analyses on the relationship between the newly derived indicators on firms' economic competencies on the one hand and firm performance on the other (e.g., through productivity analysis) would provide additional insight into the validity of these indicators. For this purpose, time-series data on both platform-based indicators and firm performance measures would be required.

8 References

Adarov, Amat; Stehrer, Robert (2019): Tangible and Intangible Assets in the Growth Performance of the EU, Japan and the US. In *wiiw Research Report* No. 442.

Aguado, David; Andrés, José C.; García Izquierdo, Antonio León; Rodríguez, Jesús (2019): LinkedIn "Big Four": Job Performance Validation in the ICT Sector. In *Journal Of Work And Organizational Psychology-Revista de Psicología del Trabajo y de las Organizaciones* 35 (2), pp. 53–64.

Arora, Amit; Arora, Anshu Saxena; Palvia, Shailendra (2014): Social Media Index Valuation: Impact of Technological, Social, Economic, and Ethical Dimension. In *Journal of Promotion Management* 20 (3), pp. 328–344.

Awano, Gaganan; Franklin, Mark; Haskel, Jonathan; Kastrinaki, Zafeira (2010a): Investing in innovation: Findings from the UK Investment in Intangible Asset Survey. In *NESTA Index Report (July)*.

Awano, Gaganan; Franklin, Mark; Haskel, Jonathan; Kastrinaki, Zafeira (2010b): Measuring investment in intangible assets in the UK: results from a new survey. In *Economic & Labour Market Review* 4 (7), pp. 66–71.

Axenbeck, Janna; Breithaupt, Patrick (2019): Web-Based Innovation Indicators–Which Firm Website Characteristics Relate to Firm-Level Innovation Activity? In *ZEW-Centre for European Economic Research Discussion Paper* (19-063).

Banerji, Devika; Reimer, Torsten (2019): Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful? In *Computers in Human Behavior* 90, pp. 46–52.

Behrens, Vanessa; Berger, Marius; Hud, Martin; Hünermund, Paul; Iferd, Younes; Peters, Bettina; Rammer, Christian; Schubert, Torben (2017): Innovation activities of firms in Germany-Results of the German CIS 2012 and 2014: Background report on the surveys of the Mannheim Innovation Panel Conducted in the Years 2013 to 2016. ZEW-Leibniz Centre for European Economic Research.

Bertschek, Irene; Kesler, Reinhold (2018): Let the User Speak: Is Feedback on Facebook a Source of Firms' Innovation? ZEW - Leibniz Centre for European Economic Research (17-015). Available online at https://ideas.repec.org/p/zbw/zewdip/17015.html.

Bruhn, Manfred; Schoenmueller, Verena; Schäfer, Daniela B. (2012): Are Social Media Replacing Traditional Media in Terms of Brand Equity Creation? In *Management Research Review* 35 (9), pp. 770–790.

Chen, Wen (2018): Cross‐Country Income Differences Revisited: Accounting for the Role of Intangible Capital. In *Review of Income and Wealth* 64 (3), pp. 626–648.

Chen, Wen; Niebel, Thomas; Saam, Marianne (2016): Are Intangibles More Productive in ICT-Intensive Industries? Evidence from EU Countries. In *Telecommunications Policy* 40 (5), pp. 471–484.

Chiang, Johannes Kuo-Huie; Suen, Hung-Yue (2015): Self-Presentation and Hiring Recommendations in Online Communities: Lessons From LinkedIn. In *Computers in Human Behavior* 48, pp. 516–524.

Choi, Hyunyoung; Varian, Hal (2012): Predicting the Present with Google Trends. In *Economic record* 88, pp. 2–9.

Chollet, Franois (2015): Keras: GitHub. Available online at https://github.com/fchollet/keras.

Chung, Sunghun; Animesh, Animesh; Han, Kunsoo; Pinsonneault, Alain (2015): The Business Value of Firms' Social Media Efforts: Evidence from Facebook. In : Proceedings of the 17th International Conference on Electronic Commerce 2015, pp. 1–8.

Claussen, Jörg; Peukert, Christian (2019): Obtaining Data from the Internet: A Guide to Data Crawling in Management Research. In *Available at SSRN 3403799*.

Corrado, Carol; Haltiwanger, John; Sichel, Daniel (2005): Measuring Capital and Technology: An Expanded Framework. In Carol Corrado, John Haltiwanger, Daniel Sichel (Eds.): Measuring Capital in the New Economy: University of Chicago Press, pp. 11–46.

Corrado, Carol; Haltiwanger, John; Sichel, Daniel (2009): Intangible Capital and US Economic Growth. In *Review of Income and Wealth* 55 (3), pp. 661–685.

Corrado, Carol; Haskel, Jonathan; Jona-Lasinio, Cecilia; Iommi, Massimiliano (2013): Innovation and Intangible Investment in Europe, Japan and the United States. In *Oxford Review of Economic Policy* 29 (2), pp. 261–286.

Corrado, Carol; Haskel, Jonathan; Jona‐Lasinio, Cecilia (2017): Knowledge Spillovers, ICT and Productivity Growth. In *Oxford Bulletin of Economics and Statistics* 79 (4), pp. 592–618.

Coursaris, Constantinos K.; van Osch, Wietske; Balogh, Brigitte A. (2016): Do Facebook Likes Lead to Shares or Sales? Exploring the Empirical Links between Social Media Content, Brand Equity, Purchase Intention, and Engagement. In : Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS), pp. 3546–3555.

Crass, Dirk; Licht, Georg; Peters, Bettina (2014): Intangible Assets and Investments at the Sector Level - Empirical Evidence for Germany. In A. Bounfour, T. Miyagawa (Eds.): Intangibles, Market Failure and Innovation Performance: Springer International Publishing.

Davis, Donald R.; Dingel, Jonathan I.; Monras, Joan; Morales, Eduardo (2019): How Segregated Is Urban Consumption? In *Journal of political economy* 127 (4), pp. 1684–1738.

Di Ubaldo, Mattia; Siedschlag, Iulia (2020): Investment in Knowledge‐Based Capital and Productivity: Firm‐Level Evidence from a Small Open Economy. In *Review of Income and Wealth*. DOI: 10.1111/roiw.12464.

Dong, Lei; Ratti, Carlo; Zheng, Siqi (2019): Predicting Neighborhoods' Socioeconomic Attributes Using Restaurant Data. In *Proceedings of the National Academy of Sciences* 116 (31), pp. 15447–15452.

European Commission (2014): Flash Eurobarometer 369 (Investing in Intangibles: Economic Assets and Innovation Drivers for Growth): European Commission, Luxembourg. Available online at https://doi.org/10.4232/1.11908.

Friedman, Jerome; Hastie, Trevor; Tibshirani, Robert (2001): The Elements of Statistical Learning: Springer series in statistics New York (10).

Gentzkow, Matthew; Kelly, Bryan; Taddy, Matt (2019): Text as Data. In *Journal of Economic Literature* 57 (3), pp. 535–574.

Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; Brilliant, Larry (2009): Detecting Influenza Epidemics Using Search Engine Query Data. In *Nature* 457 (7232), pp. 1012–1014.

Glaeser, Edward L.; Kim, Hyunjin; Luca, Michael (2018): Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. In *AEA Papers and Proceedings* 108, pp. 77–82. DOI: 10.1257/pandp.20181034.

Gök, Abdullah; Waterworth, Alec; Shapira, Philip (2015): Use of Web Mining in Studying Innovation. In *Scientometrics* 102 (1), pp. 653–671.

Gortmaker, Jeff; Jeffers, Jessica; Lee, Michael (2020): Labor Reactions to Financial Distress: Evidence from LinkedIn Activity. In *Available at SSRN*. DOI: 10.2139/ssrn.3456285.

Haskel, Jonathan; Westlake, Stian (2018): Capitalism without Capital: The Rise of the Intangible Economy: Princeton University Press.

Ji, Yuan; Rozenbaum, Oded; Welch, Kyle (2017): Corporate Culture and Financial Reporting Risk: Looking Through the Glassdoor. In *Available at SSRN*. DOI: 10.2139/ssrn.2945745.

Kinne, Jan; Lenz, David (2019): Predicting Innovative Firms Using Web Mining and Deep Learning. In *ZEW-Centre for European Economic Research Discussion Paper* (19-01).

Krüger, Miriam; Kinne, Jan; Lenz, David; Resch, Bernd (2020): The Digital Layer: How Innovative Firms Relate on the Web. In *ZEW-Centre for European Economic Research Discussion Paper* (20-003).

Lucas, Robert (1988): On the Mechanics of Economic Development. In *Journal of Monetary Economics* 22 (1), pp. 3–42.

Luo, Xueming; Zhang, Jie; Duan, Wenjing (2013): Social Media and Firm Equity Value. In *Information Systems Research* 24 (1), pp. 146–163.

Misirlis, Nikolaos; Vlachopoulou, Maro (2018): Social Media Metrics and Analytics in Marketing–S3M: A Mapping Literature Review. In *International Journal of Information Management* 38 (1), pp. 270–276.

Niebel, Thomas; O'Mahony, Mary; Saam, Marianne (2017): The Contribution of Intangible Assets to Sectoral Productivity Growth in the EU. In *Review of Income and Wealth* 63, S49-S67.

Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent (2011): Scikit-learn: Machine learning in Python.

Perani, Giulio; Guerrazzi, Marco (2012): The Statistical Measurement of Intangible Assets: Methodological Implications of the Results of the ISFOL 2011 Pilot Survey. mimeo (available upon request from the authors).

Peters, Bettina; Rammer, Christian (2013): Innovation Panel Surveys in Germany. In Fred Gault (Ed.): Handbook of Innovation Indicators and Measurement: Edward Elgar Publishing, pp. 135–177.

Pisano, Sabrina; Lepore, Luigi; Lamboglia, Rita (2017): Corporate Disclosure of Human Capital via LinkedIn and Ownership Structure. In *Journal of Intellectual Capital* 18 (1), pp. 102–127.

Pukelis, Lukas; Stanciauskas, Vilius (2019): Using Internet Data to Compliment Traditional Innovation Indicators. Paper presented at the International Conference on Public Policy (ICPP4). Available online at https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf.

Rammer, Christian; Roth, Felix; Trunschke, Markus (2020): Measuring Organisation Capital at the Firm Level: A Production Function Approach. In *ZEW-Centre for European Economic Research Discussion Paper* (20-021).

Romer, Paul M. (1986): Increasing Returns and Long-Run Growth. In *Journal of political economy* 94 (5), pp. 1002–1037.

Romer, Paul M. (1990): Endogenous Technological Change. In *Journal of political economy* 98 (5, Part 2), S71-S102.

Roth, Felix (2019): Intangible Capital and Labour Productivity Growth: A Review of the Literature. Hamburg Discussion Papers in International Economics.

Roth, Felix; Thum, Anna-Elisabeth (2013): Intangible Capital and Labor Productivity Growth: Panel Evidence for the EU from 1998-2005. In *Review of Income and Wealth* 59 (3), pp. 486–508.

Tirunillai, Seshadri; Tellis, Gerard J. (2012): Does Online Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance. In *Marketing Science* 31 (2), pp. 198–215.

Zide, Julie; Elman, Ben; Shahani-Denning, Comila (2014): LinkedIn and Recruitment: How Profiles Differ Across Occupations. In *Employee Relations* 36 (5), pp. 583–604.

Appendix

A Additional Tables

Table A-1: Overview of Possible Dimensions on Platforms

	Kununu	**Facebook**
Information on the start page	Grade, scale, recommendations, hits, benefits	Number of "likes" obtained without the Facebook Graph API.
Historic information	Individual/detailed ratings (including "company image" and "on-the-job training/career development") are available with a time stamp, so the data can be reconstructed historically.	Get company and user contributions including comments and "likes" via the Facebook Graph API. You only get the current number of fans.
Remarks	Problem of deletion/change	API access massively restricted in the wake of the Facebook–Cambridge Analytica data scandal early 2018.

Table A-2: Summary Statistics - Kununu - Full Sample

	N	Mean	Median	SD	Min	Max
Training: Kununu rating	813	3.29	3.35	0.77	1	5
Image: Kununu rating	805	3.56	3.67	0.79	1	5
Training expenditures (MEUR)	1568	0.42	0.019	7.96	0	300
Marketing expenditures (MEUR)	1548	2.50	0.040	46.0	0	1480
Turnover (MEUR)	1961	252.3	10.3	2222.4	0	57550
Number of employees	2051	774.8	70	6643.5	0	156487
Number of Kununu ratings (Training)	2114	9.91	2	43.3	0	1174
Number of Kununu ratings (Image)	2114	9.83	2	43.1	0	1171
ln(Training: Kununu rating)	813	1.16	1.21	0.26	0	1.61
ln(Image: Kununu rating)	805	1.24	1.30	0.25	0	1.61
ln(Training expenditures)	1356	-3.49	-3.69	1.84	-7.71	5.70
ln(Marketing expenditures)	1321	-2.63	-2.81	2.19	-8.11	7.30
ln(Turnover)	1957	2.47	2.33	2.15	-5.30	11.0
ln(Number of employees)	2045	4.35	4.25	1.74	0	12.0

Note: Full Sample means the merge of the MIP 2017 survey with the identified Kununu profiles (see Figure 3-2). The "Number of Kununu ratings" shows the descriptive statistics for all 2,114 MIP 2017 firms with a Kununu profile. The number of observations for "Kununu rating" is lower as we restrict the data to firms with at least 4 ratings between January 2017 and August 2018.

Table A-3: Summary Statistics - Facebook - Full Sample

	N	Mean	Median	SD	Min	Max
Image: Facebook likes	1498	15905.7	215.5	142384.2	1	3687320
Marketing expenditures (MEUR)	1165	1.95	0.020	44.0	0	1480
Turnover (MEUR)	1425	134.5	4.23	1891.5	0	57550
Number of employees	1498	465.2	35	5436.4	0	156487
ln(Image: Facebook likes)	1498	5.71	5.37	2.30	0	15.1
ln(Marketing expenditures)	1012	-3.24	-3.51	2.11	-8.11	7.30
ln(Turnover)	1421	1.69	1.46	2.09	-5.30	11.0
ln(Number of employees)	1492	3.77	3.56	1.69	0	12.0

Note: Full Sample means the merge of the MIP 2017 survey with the identified Facebook profiles (see Figure 3-2).

Table A-4: Sector Coverage Kununu

	MIP 2017		**Full Sample**		**Estimation Sample (Training)**		**Estimation Sample (Image)**	
	N	Percent	N	Percent	N	Percent	N	Percent
A - Agriculture, forestry and fishing	13	0.16						
B - Mining and quarrying	98	1.18	5	0.24	1	0.19	2	0.41
C - Manufacturing	3664	44.26	996	47.11	241	46.44	220	44.72
D - Electricity, gas, steam, air conditioning supply	135	1.63	56	2.65	11	2.12	9	1.83
E - Water supply,sewerage,waste management,remediation	386	4.66	44	2.08	11	2.12	8	1.63
F - Construction	204	2.46	25	1.18	4	0.77	5	1.02
G - Wholesale, retail trade, repair of motor vehicles	435	5.25	106	5.01	26	5.01	27	5.49
H - Transportation and storage	543	6.56	109	5.16	20	3.85	16	3.25
I - Accommodation and food service activities	15	0.18	1	0.05				
J - Information and communication	615	7.43	260	12.30	78	15.03	79	16.06
K - Financial and insurance activities	255	3.08	95	4.49	23	4.43	21	4.27
L - Real estate activities	53	0.64	7	0.33				
M - Professional, scientific and technical activities	1329	16.05	282	13.34	75	14.45	75	15.24
N - Administrative and support service activities	502	6.06	119	5.63	28	5.39	28	5.69
O - Public administration and defence, compulsory social security	2	0.02						
P - Education	10	0.12	4	0.19				
Q - Human health and social work activities	2	0.02						
R - Arts, entertainment and recreation	7	0.08	3	0.14	1	0.19	1	0.20
S - Other service activities	10	0.12	2	0.09			1	0.20
Total	8278	100.00	2114	100.00	519	100.00	492	100.00

Table A-5: Size Classes Kununu

# of Employees	MIP 2017		Full Sample		Estimation Sample (Training)		Estimation Sample (Image)	
	N	Percent	N	Percent	N	Percent	N	Percent
0-9	2326	28.69	224	10.92	21	4.05	24	4.88
10-49	3274	40.38	633	30.86	108	20.81	108	21.95
50-249	1752	21.61	708	34.52	190	36.61	176	35.77
250+	755	9.31	486	23.70	200	38.54	184	37.40
Total	8107	100.00	2051	100.00	519	100.00	492	100.00

Note: For 171 firms in the MIP 2017 sample, the number of employees is not available in the data.

Table A-6: Sector Coverage Facebook

	MIP 2017		Full Sample		Estimation Sample	
	N	Percent	N	Percent	N	Percent
A - Agriculture, forestry and fishing	13	0.16	1	0.06		
B - Mining and quarrying	98	1.18	9	0.58	3	0.32
C - Manufacturing	3664	44.26	686	44.57	407	43.11
D - Electricity, gas, steam, air conditioning supply	135	1.63	28	1.82	19	2.01
E - Water supply, sewerage, waste management, remediation	386	4.66	37	2.40	24	2.54
F - Construction	204	2.46	27	1.75	19	2.01
G - Wholesale, retail trade, repair of motor vehicles	435	5.25	97	6.30	66	6.99
H - Transportation and storage	543	6.56	104	6.76	66	6.99
I - Accommodation and food service activities	15	0.18	5	0.32	4	0.42
J - Information and communication	615	7.43	173	11.24	112	11.86
K - Financial and insurance activities	255	3.08	57	3.70	32	3.39
L - Real estate activities	53	0.64	7	0.45	5	0.53
M - Professional, scientific and technical activities	1329	16.05	191	12.41	116	12.29
N - Administrative and support service activities	502	6.06	109	7.08	67	7.10
O - Public administration and defence, compulsory social security	2	0.02				
P - Education	10	0.12	2	0.13	1	0.11
Q - Human health and social work activities	2	0.02	1	0.06	1	0.11
R - Arts, entertainment and recreation	7	0.08	3	0.19	2	0.21
S - Other service activities	10	0.12	2	0.13		
Total	8278	100.00	1539	100.00	944	100.00

Table A-7: Size Classes Facebook

	MIP 2017		Full Sample		Estimation Sample	
# of Employees	N	Percent	N	Percent	N	Percent
0-9	2326	28.69	277	18.49	132	13.98
10-49	3274	40.38	573	38.25	379	40.15
50-249	1752	21.61	430	28.70	292	30.93
250+	755	9.31	218	14.55	141	14.94
Total	8107	100.00	1498	100.00	944	100.00

Note: For 171 firms in the MIP 2017 sample, the number of employees is not available in the data.

Table A-8: Robustness Check: OLS Regressions Kununu - A Least 3 Ratings

Dependent Variable:	(1) Training: Kununu rating	(2) ln(Training: Kununu rating)	(3) Image: Kununu rating	(4) ln(Image: Kununu rating)
ln(Training expenditures)	0.0945***	0.0341***		
	(2.70)	(2.87)		
ln(Marketing expenditures)			0.0886***	0.0274***
			(3.37)	(3.17)
ln(Turnover)	0.0199	0.0119	-0.0403	-0.00998
	(0.47)	(0.81)	(-0.88)	(-0.68)
ln(Number of employees)	-0.0916	-0.0331*	-0.0544	-0.0160
	(-1.64)	(-1.71)	(-0.99)	(-0.88)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.126	0.129	0.136	0.137
Observations	613	613	582	582

Robust t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table A-9: Robustness Check: OLS Regressions Kununu - A Least 5 Ratings

Dependent Variable:	(1) Training: Kununu rating	(2) ln(Training: Kununu rating)	(3) Image: Kununu rating	(4) ln(Image: Kununu rating)
ln(Training expenditures)	0.0820**	0.0284**		
	(2.00)	(2.10)		
ln(Marketing expenditures)			0.0907***	0.0275***
			(3.11)	(3.05)
ln(Turnover)	0.00118	0.00684	-0.0714	-0.0196
	(0.03)	(0.46)	(-1.39)	(-1.20)
ln(Number of employees)	-0.0659	-0.0266	-0.0211	-0.00571
	(-1.15)	(-1.38)	(-0.34)	(-0.28)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.131	0.131	0.105	0.104
Observations	433	433	413	413

Robust t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

B Additional Graphs

Figure B-1: Histogram Training: Kununu Rating

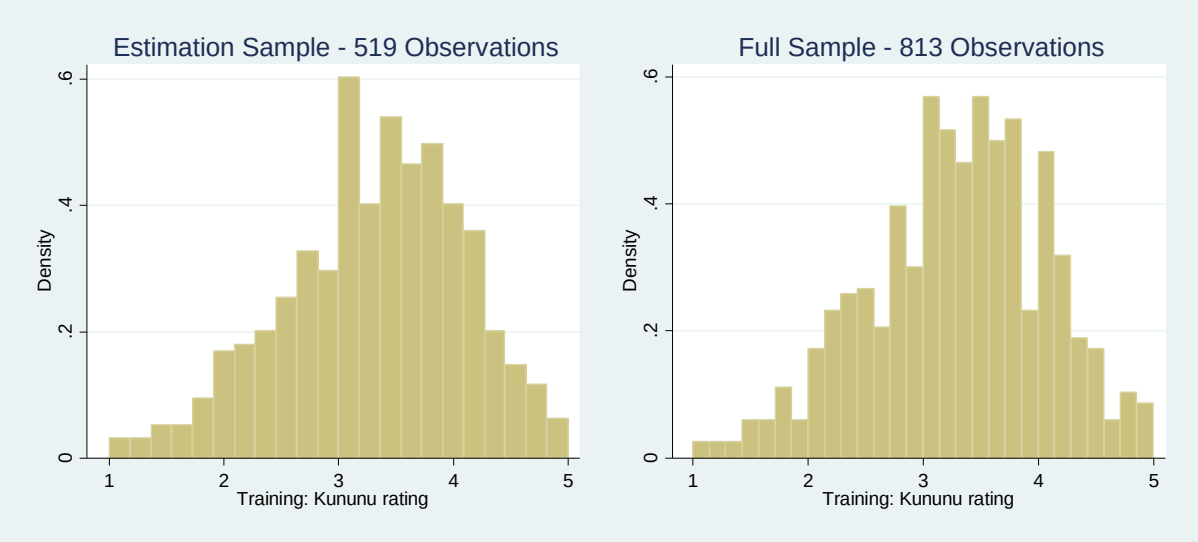


Figure B-2: Histogram Image: Kununu Rating

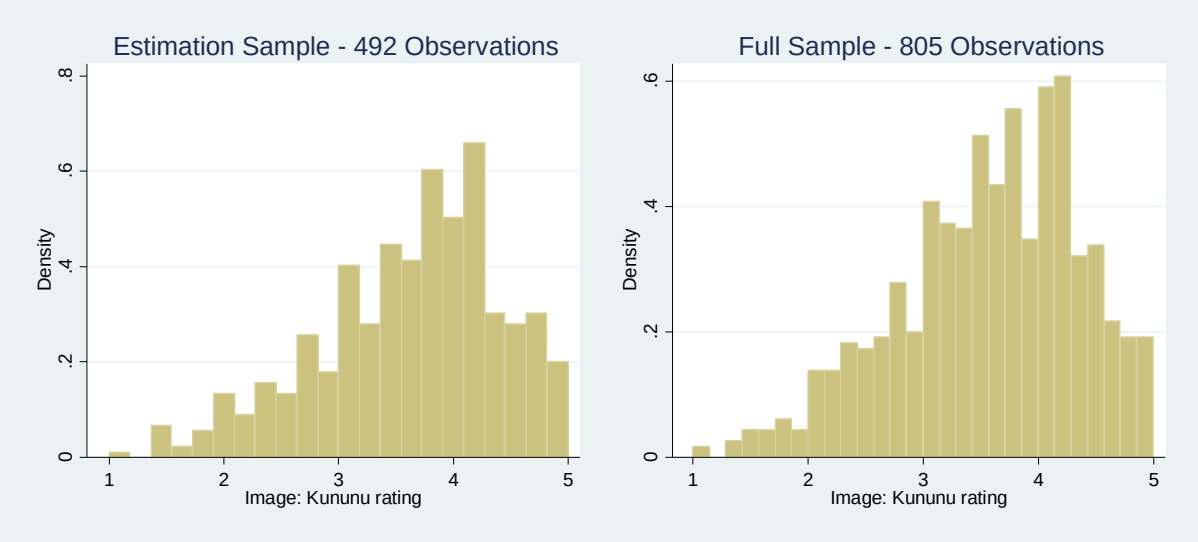


Figure B-3: Histogram Ln(Image: Facebook Likes)

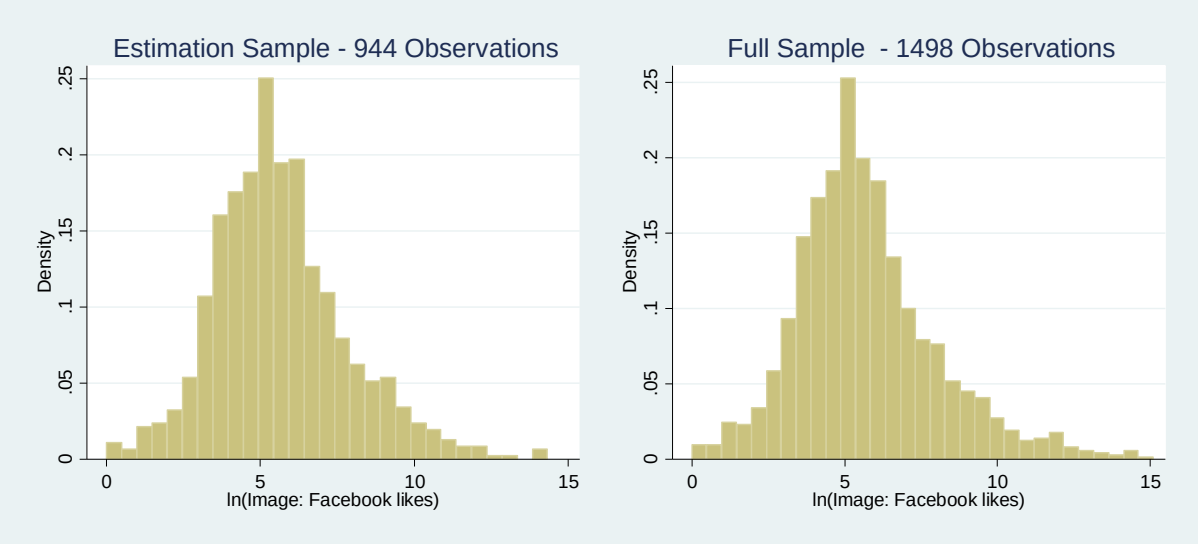


Figure B-4: Scatterplot Training: Kununu Rating vs MIP Ln(Training Expenditures)

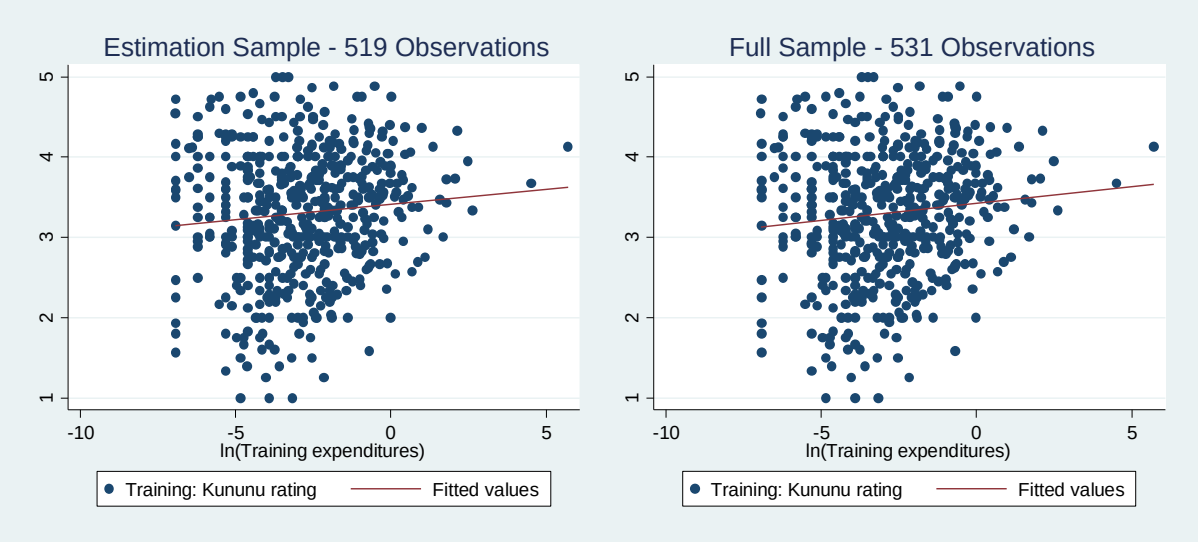


Figure B-5: Scatterplot Image: Kununu Rating vs MIP Ln(Marketing Expenditures)

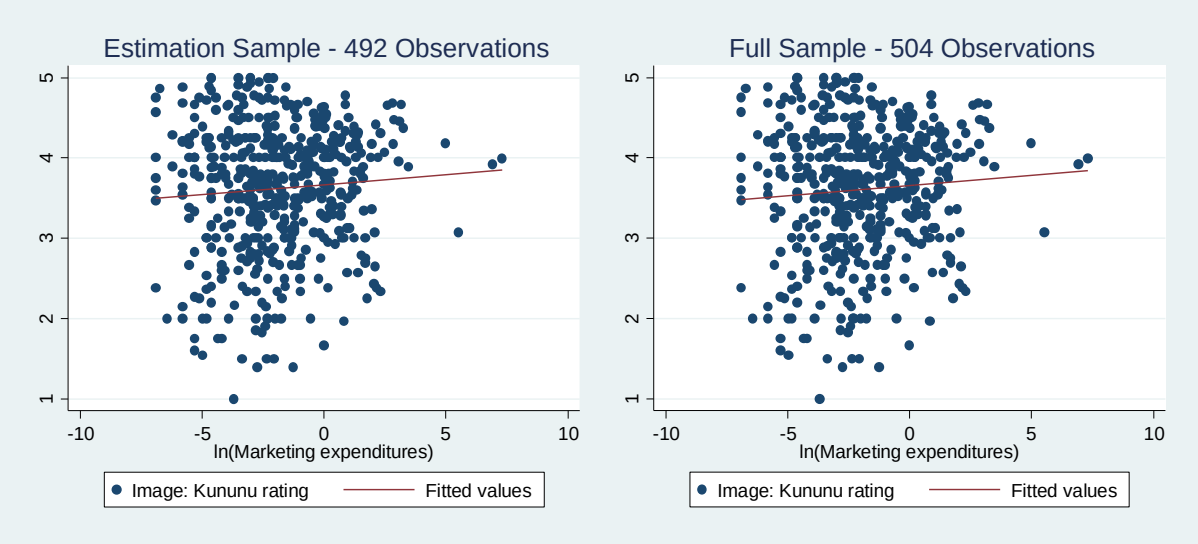


Figure B-6: Scatterplot: Ln(Facebook Likes) vs MIP Ln(Marketing Expenditures)