

Predicting Startup Survival from Digital Traces: Towards a Procedure for Early Stage Investors

Short Paper

Torben Antretter

University of St. Gallen
Institute of Technology Management
St. Gallen, Switzerland
torben.antretter@unisg.ch

Ivo Blohm

University of St. Gallen
Institute of Information Management
St. Gallen, Switzerland
ivo.blohm@unisg.ch

Dietmar Grichnik

University of St. Gallen
Institute of Technology Management
St. Gallen, Switzerland
dietmar.grichnik@unisg.ch

Abstract

We investigate whether digital traces can be used to predict early stage startup survival. Based on common survival factors from the entrepreneurship literature, we mined the digital footprints of 542 entrepreneurs and their ventures. Using a context-specific text mining approach, we performed a bootstrapping simulation in which we predict 5-year survival for different survival rates that range from 50% to 10%. Our results indicate that we can predict 5-year survival with an accuracy of up to 91%. With this study, we will provide an evidence-based taxonomy of digital traces for predicting early stage startup survival, identify the most important digital traces for doing so and benchmark our predictive approach against the actual investments of 339 business angels.

Keywords: business angels, survival prediction, machine learning, digital traces, social media

Introduction

Startups are an essential factor to economic development. A study from the United States showed that between 1977 and 2005, startups produced – on average – 3 million new jobs every year. In the same period, established companies cut back on over 1 million jobs annually (Kane 2010). To set up their businesses and grow, most startups require equity financing by early stage investors – business angels (BAs). BAs are private investors that usually invest at early stages of development – often before the feasibility of a new product or market is proven (Morrisette 2007). Therefore, they often lack reliable data sources to assess the quality of a startup and to determine the risk of their investments. Existing work has helped us understand their investment criteria, but has neglected the fact that up to 78% of their portfolio investments fail to survive for five years in the marketplace. This downside risk leads even the most experienced BAs to lose money on over half of their deals (Song et al. 2008).

There have been few attempts to build machine learning algorithms to improve investment decisions into startups (e.g., Hunter and Zaman 2017; Krishna et al. 2016). These studies provide promising results regarding the predictability of entrepreneurial outcomes. However, all present findings result from transactions in the venture capital (VC) field, where decisions are made at later stages – after a startup has proven the validity of its business model. Machine learning approaches to assist BA investment decisions are lacking as machines fail to process *soft* types of information and to make intuitive judgments (Petersen

2004). In these situations, humans are still considered as being the *gold standard* in decision-making. However, BAs' investment choices underlie the limitations of their bounded rationality. These cognitive limits often lead to biased judgments and may result in erroneous predictions. Hence, identifying relevant data sources to quantify soft signals for BA investing to build machine learning algorithms that are not constraint by limited cognitive resources has tremendous potential for BAs and the economy.

A growing proportion of human activities leave digital traces on the Internet (Lambiotte and Kosinski 2014). This data offers unprecedented insights into behavioral patterns and characteristics of individuals. Research has shown that these digital footprints can be transformed into machine-readable soft signals, such as personality traits (Quercia et al. 2012). Entrepreneurs make use of social media such as *LinkedIn* or *Twitter* to maintain relationships, communicate with customers and publish key milestones of their businesses (Yang and Berger 2017). However, not only social media but also platforms such as *Crunchbase* provide valuable insights into the structures and achievements of early stage startups. The challenge of utilizing digital traces is that they are inherently messy. The information is self-reported and often does not undergo any review process. Taking into consideration the practical difficulties of BAs to choose among several investment opportunities and the increasing size of our digital footprints, the question arises: How can digital traces be utilized to predict startup survival and assist early stage investment decisions?

Our study aims to mine digital traces of entrepreneurs in a way to build predictive models for early stage startup survival. Further, it will provide a sensitivity analysis to test whether portfolio selections based on survival probability can be extrapolated to achieve higher investment returns. We chose survival as it is one of the essential indicators for entrepreneurial success (Wennberg et al. 2010). We use a combination of different data sources but mainly explore the precious data that resides within social media activities of early stage startups and their founders. We employ a text mining procedure to automate the evaluation of social media profiles based on common startup survival factors. Our research-in-progress shows that we can predict 5-year survival with an accuracy of up to 91%. Our results may contribute to several streams of information systems (IS) and entrepreneurship research. First, we conceptualize empirical findings from the startup survival literature and provide an evidence-based taxonomy of digital traces that can be used to predict early stage startup survival and will identify a manageable set of the most important digital traces as survival predictors. Second, we will compare our predictive approach against the investments that have been performed by 339 business angels. Third, our study will expand current IS research on the business value of collecting, aggregating and combining publicly available, unstructured and messy digital traces.

Related Work

Predictors of Early Stage Startup Survival

Given the high failure rates (up to 78%), explaining startup survival has attracted much scholarly attention. Although current evidence is controversial and fragmented (Delmar and Shane 2006), research has been mainly concerned with studying survival factors on the individual, organizational and environmental level. *Individual factors* refer to traits of the founder, i.e., personality (Ciavarella et al. 2004), functional background and formal education (Delmar and Shane 2006 for a review) as well as social networks (Raz and Gloor 2007). In line with this research and given the fact that many startups are online or technology-based businesses, prior technical and entrepreneurial experiences may also be relevant, because similar business experiences help entrepreneurs to acquire resources, attract customers and hire the right employees (Mayer-Haug et al. 2013). *Organizational factors* refer to the characteristics of the startup itself. Factors that have been positively associated with survival include the legal form (Wennberg et al. 2010), team size, financial resources (Headd 2003), industry type and product innovativeness (Hyytinen et al. 2015). *Environmental factors* refer to external influences that affect survival. For instance, market perception of the product and demand growth are positively related to survival (Colombelli et al. 2013). Audretsch and Mahmood (1994) show that higher competitiveness leads to lower survival rates, while being located in *startup hotspots* has a positive effect (Pe'er and Keil 2013).

Machine Learning as Possible Approach for Predicting Startup Survival

Predicting future events through computational statistical methods has become much more accessible in recent years due to advancements in machine learning and artificial intelligence. Predictive modeling has been used to make highly accurate predictions in the context of stock markets or bankruptcy of established

organizations (e.g., Creamer et al. 2016). Given appropriate data, machine intelligence may be unbiased, free of social or affective contingency and does not suffer from cognitive resources limitations (Blattberg and Hoch 1990). Research from the VC field shows promising results in predicting the outcome of startups. For instance, Krishna et al. (2016) used financial metrics, such as the number of investment rounds, to accurately predict the outcome of 11,000 companies using Crunchbase. Predicting the success of VC-backed startups, Hunter and Zaman (2017) provide a useful theoretical framework for picking the most successful investment opportunities. Using random draws, the authors constructed portfolios and showed that the exit rates of those portfolios are as high as 60% and thus much higher than exit rates in conventional VC portfolios (Hochberg et al. 2007). However, these studies use data on later stage investments where startups have already received several financing rounds. These data points do not exist when BAs invest and similar research for early stage investing is scarce. Therefore, the goal of this study is to identify data sources that may be able to approximate known factors that influence new venture survival. For instance, product perception, which can, for later stage ventures be measured through product revenues, is not an available metric for early stage ventures. However, the public's reactions to a product on social web may serve as an approximation to how well a product is received in the market (see also Mollick 2016). Most notably for the purpose of this study, Dellermann et al. (2017) propose a hybrid model of machine learning and collective intelligence to capture the soft signals of BA investment decisions. This short review shows that initial attempts to predict the outcome of startups are promising. However, as machine learning approaches can only process available data, prior research has either focused on organizational and environmental factors for later stage startups with proven business models or proposed workarounds such as collective expert intelligence to capture the necessary soft signals in BA investing. Thus, systematically utilizing digital traces may reflect a valid alternative for early stage investing that existing research has not yet considered.

Digital Traces as a Data Source to Predict Early Stage Startup Survival

Previous work that leveraged digital traces to predict future outcomes has suggested that this data may provide ways to capture all three categories of startup survival factors (see Table 1).

Table 1. Digital Traces and Survival Factors

Dimension	Factor	a)	b)	Factor proxy	Source of trace	Selected reference
Individual Factors	Personality		✓	• Big5 personality traits • Online communication	• Website / LinkedIn • Twitter	Quercia et al. (2012)
	Mgmt. experience		✓	• Managerial job positions	• LinkedIn / Website	Hunter and Zaman (2017)
	Startup experience		✓	• Prior startup projects	• LinkedIn / Website	
	Tech. experience		✓	• Technical job positions • Content on tech-blogs	• LinkedIn • Website / GitHub	
	Formal education	✓		• Highest degree	• LinkedIn / Website	Gloor et al. (2013)
	Network structure		✓	• Connections/followers • Network centrality	• LinkedIn • Twitter	
Organizational Factors	Legal form	✓		• Type of legal form	• Commercial register	Muller et al. (2016)
	Team size	✓		• No. of co-founders • Number of females	• Website • Commercial register	Hunter and Zaman (2017)
	Financial resources	✓		• Crowdfunding • Financing rounds	• Commercial register • Crunchbase	Mollick and Kuppawamy (2014)
	Industry type	✓		• Startup industry	• Website	Böhm et al. (2017)
	B2B vs. B2C	✓		• Description • Industry tags	• Website • LinkedIn	
	Size	✓		• No. of employees • Revenue	• LinkedIn • Crunchbase	Hunter and Zaman (2017)
	Age	✓		• Founding date	• Commercial register	Brüderl et al. (1992)
	Innovativeness	✓		• No. of patents • Website content	• Patent office • Company website	Hyytinen et al. (2015)
Environmental Factors	Product perception		✓	• No. likes / Sentiments • Sentiment of posts	• Twitter • Google Trends	Gruhl et al. (2005)
	Demand growth		✓	• Growth of market	• Census data	Müller et al. (2016)
	Competitiveness		✓	• Website ranking • Number similar websites	• Similarweb • National statistics	Choi and Varian (2012)
	Startup hotspot	✓		• City location	• Startup Monitor	Pe'er and Keil (2013)
	Qual. of workforce		✓	• Workforce distribution	• National statistics	Müller et al. (2016)

a) Readily quantifiable aspects, b) Factors where intuitive judgment is needed

Individual Factors. The entrepreneurship literature suggests a strong influence of the founders' personality on survival (Ciavarella et al. 2004). Studies have demonstrated that personality can be predicted based on digital traces contained from personal websites (Marcus et al. 2006), Twitter profiles (Quercia et al. 2012) or the language used by individuals (Fast and Funder 2008). These results suggest that self-descriptive online content of early stage startups and their founders can be used to make personality predictions (Kosinski et al. 2013). Similarly, studies have used LinkedIn and company websites to generate founder information and argued that they are the most accurate external data sources for educational and career information because the information is “entered by each registered user and is periodically updated” (Ko and McKelvie 2018: 7). Hunter and Zaman (2017) argue that LinkedIn profiles often contain information on previous startup experience. Finally, a founder's activity in online social networks may be important. For instance, Gloor et al. (2013) found that the more central entrepreneurs are in the different types of online social networks, the more successful they are.

Organizational Factors. Other data sources that could be leveraged to generate information about the organizational structures of startups are commercial registers that provide data such as legal form, founding date, initial share capital and domicile address. Scholars have made use of this data source to study startup survival (e.g., Brüderl et al. 1992). Further, to study the characteristics of different business models, Böhm et al. (2017) manually searched firm's websites, press reports and interviews with founders that were publically available. Therefore, we expect that there is a high potential to leverage such digital traces to obtain machine-readable data, e.g., on business model types and industries. To further account for a products “innovativeness” we captured whether a startup could be described as a “high-tech” startup or not, as such ventures show higher degrees of innovativeness than low-tech companies (Hyytinen et al. 2015).

Environmental Factors. Product sales – and thus the public perception of a startups product – is important for startup survival (Bayus and Agarwal 2007). Research on electronic word-of-mouth suggests that the way in which online communities talk about a product may be indicative of a product's sales (Gruhl et al. 2005). For instance, Asur and Huberman (2010) used Twitter data to predict film box office revenues accurately. Similarly, Choi and Varian (2012) used the Google Trend Indicator to successfully predict car sales, claims for unemployment benefits and travel destinations. Using location data such as a startup's domicile address, scholars have linked the location to different economic variables such as competitiveness (Muller et al. 2016) or the closeness to “startup hotspots” (Pe'er and Keil 2013).

Methodology

Our goal is to leverage digital traces of entrepreneurs to predict startup survival. Following Fayyad et al. (1996), we first collected a set of individual, organizational and environmental predictors from the digital traces of startups and their founders. We then extracted, loaded and transformed them into a machine-readable format. Finally, we applied different machine learning algorithms to predict whether a startup survived or failed. In a second step, we will provide a sensitivity analysis based on real BA investment data to test whether we can also predict future startup performance.

Data Collection

Our study is based on the data of two large BA investment platforms from Switzerland. We use their deal monitoring system and their list of investment opportunities to create a database that consists of detailed information on the startup and founder level. The central database consists of the companies' names, addresses, websites, industry tags and the names and social media profiles of their founders. To enhance data quality, we performed a stepwise process of combining our data with the digital traces that the startups and their founders have left on the Internet. First, on the startup level, we used national commercial registers to enrich our database with fundamental information about each company, i.e., founding date, number of shareholders and legal form. Second, we used Crunchbase and the startup's websites to define its business focus (B2C vs. B2B) and industry sector. As a third step, we used Google Trends and Twitter to capture the public perception and communication behavior of each startup. Specifically, we obtained data on the new venture's Twitter accounts, i.e., the number of tweets, the likes that each startup received as well as their number of followers and friends. On the founder level, we used LinkedIn profiles to obtain the educational and occupational histories for the founders of each startup. The founders' profiles further provided us with data on their professional network as well as their self-reported interests and capabilities. Finally, we used the Apply Magic Sauce (AMS) engine (Kosinski et al. 2013) to obtain psychometric profiles

(“Big Five”) based on a self-descriptive text that the founders have provided when they applied for the BA investment platforms. We consolidated all founder data on the startup level by the arithmetic mean to make them comparable across companies. This process resulted in a dataset consisting of 731 startups.

Variables

Startup Survival. Our dependent variable is 5-year survival – a dummy indicating whether or not a company was active for more than 5 years. We followed Raz and Gloor (2007) and did a company search on the Internet to find out if a startup still had a website, and if so, what the company’s website said. It is important to note that we excluded companies that were founded within the last 5 to allow for studying 5-year survival and to enable comparability across the new ventures in our dataset.

Individual Level. We obtained *character traits* and psychometric profiles based on self-descriptive text samples that the founders have provided to the BA investment platforms when they applied for funding. Although the questions were standardized, the responding entrepreneurs were free in the way they wished to answer. Thus, our approach is in line with Fast & Funder (2008) about the relationship between words and personality. The authors found were the first to find that several categories of word use yielded large correlations with self- and acquaintance personality ratings and behavior. We inserted the texts into the AMS engine (Kosinski et al. 2013) – an API that can translate text into estimates of personality, intelligence, gender and age. The resulting personality profiles are based on the “Big Five” personality attributes – extraversion, emotional stability, agreeableness, conscientiousness and openness (Ciavarella et al. 2004).

We included *professional experiences* such as managerial, entrepreneurial and technical experience as sources of relevant know-how. As Top-Tier investment banks and consulting firms have selective recruitment processes, we also captured whether a founder has prior experience in these domains. To quantify prior experiences, we employed a text mining procedure to evaluate the founders’ prior positions as reported in their LinkedIn profiles. To make this unstructured data quantifiable, we developed a series of experience-specific wordlists, which allow us to extract the founders’ experiences *before* the founding date. For instance, a prior job was labeled as conveying entrepreneurial experience when it was described with words such as *Founder* or *Co-Founder*. Triangulating this classification with the reported duration of the position, we could quantify the number of relevant positions (jobs) as well as their duration in months for each founder. We tested the reliability of this approach by manually coding the profiles of the founders of 100 startups and calculating Cronbach Alpha as a measure of intercoder reliability. The values were above 0.7 for all types of experience. We aggregated all experiences with the arithmetic mean on the startup level.

We measured the founders’ *formal education* in four ordinal categories, i.e., Ph.D./MBA, Master’s degree, Bachelor’s degree or no college degree. Similar to measuring the founders’ professional experience, we employed a text mining procedure to evaluate the LinkedIn profiles. Further, we differentiated between technical and business education and whether a founder was enrolled in either a leading technical or business university (according to the Times Higher Education Ranking). The *size of the professional network* was measured by the average number of LinkedIn contacts that all founders had. All measures at the individual level were cut off at the date at which the focal company was founded. This approach allows us to compare the experiences of founders whose companies were established at different points in time.

Organizational Factors. To obtain the startups’ *legal form*, we used each country’s commercial register. We measured the *founding team size* by counting the number of co-founders that the startups self-reported to the BA investment platform. To account for the *business model focus*, we evaluated the industry tags provided by our research partner. When no industry tag was present that differentiated the business between a B2C and B2B focus, we did a company search on Crunchbase to determine the company’s business focus (Böhm et al. 2017). With the same process, we measured each startups industry field. We used startup sectors descriptions (e.g., FinTech or MedTech) and matched these descriptions with a *Google Trend Indicator* (GTI) at the time of founding that reflects the global search volume for each search string at that time. This process allowed us to proxy public interest in each industry. All GTI values were collected at the time of founding the venture to increase comparability across the sample.

Environmental Factors. We measured the product perception of each startup using the Twitter Search API. First, we followed Asur and Huberman (2010) and extracted data on the number of tweets to proxy the communication behavior. In a second step, we then used the number of followers, friends, and likes per tweet to account for the public perception of each startups product or service.

Predicting Startup Survival

We predict 5-year survival as this equals a typical investment horizon of BAs (Prowse 1998). However, this choice has significantly narrowed down our data set. In total, we could infer 5-year survival for 238 of the startups, while 59 startups did not survive 5 years. The other companies have been founded within the last 5 years such that we could not include them. Existing studies regarding 5-year survival of startups have reported different survival rates that range from 71% (Mudambi and Zahra 2007) to 22% (Song et al. 2008). Our sample showed a 5-year survival rate of 80%. In order to simulate the potential of digital traces for predicting startup survival for different survival rates and to mitigate potential sampling biases from our data collection procedure, we apply a stratified bootstrapping approach (Bickel and Freedman 1984). Different bootstrap simulations have been reported in information systems research and generally showed high predictive validity (Goodhue et al. 2007; Riedl et al. 2013). We defined survival rates that ranged from 50% to 10%. For each rate, we created 1000 bootstrap samples in which we have drawn 150 startups randomly with replacement from the original data. We stratified individual bootstrap samples according to the survival rates, e.g., for a 50% percent survival rate, and drew 75 companies that survived 5 years and 75 companies that did not survive 5 years. Within each bootstrap sample, we built a binary classification model using the sampled data as training data. In addition, we created stratified testing data for each bootstrap iteration from the non-sampled data that accounted for 30% of instances of the training set and that mirrored the percentage of surviving companies as used in the training data. Thus, bootstrapped training and data sets were equally (un-) balanced in each bootstrapping iteration. We constructed several classifiers using and randomForest or gradient boosting (Friedman 2001). In order to measure the performance of our approach, we calculated several measures on the testing data sets. For overall performance we calculated the accuracy of our prediction (the percentage of startups that were correctly classified as “dead” or “alive”) as well as the Area Under the Curve (AUC). Further, we calculated the recall/sensitivity (the percentage of survived startups that have been correctly classified as “alive” from all survived startups) and precision (the percentage of startups that were classified as “alive” from all startups that were classified as “alive”). In order to create the final predictions, we followed the ideas of Riedl et al. (2013) and aggregated the predictions of the single bootstrap iterations by arithmetic mean. In total, both classification algorithms performed quite similarly such that we only report results of gradient boosting using the “xgboost” implementation for R. We used 200 trees with a maximal depth of 10 nodes and a learning rate of 0.1 (eta).

Planned Sensitivity Analysis: Predicting Startup Performance

We aim to test the relevance of our procedure by comparing our results with BA portfolios from our research partners. We have data on financial valuations for 66 startups in which the 339 BAs have invested. Thus, we will extend our bootstrapping approach and construct hypothetical investment portfolios based on the predicted survival probabilities, e.g., such a portfolio could include the three startups with the highest survival probability (out of the 66 startups). These portfolios can then be compared against actual portfolios of the same size in which the BAs of our research partners have invested. Based on the valuation data, we can then compare the internal rates of return from our hypothetical portfolio with those of actual BAs.

Preliminary Results

We report our results in Figure 1 and 2. Figure 1 depicts the results from the aggregated predictions across all bootstrap samples. Our results show that the analysis provides quite accurate results across all rates of 5-year survival. On average, prediction accuracies range between 91% (50% survival rate) and 55% (10% survival rate). The AUC values indicate that our model performs best for the 50% survival rate and that model the performance diminishes with decreasing survival rates. The highest AUC is reached for a survival rate of 50% where it reaches 0.78. Across all survival rates recall and precision are high. For instance, in the realistic case of 20% startup survival (Song et al. 2008) our model shows a recall of 77% and precision of 84%. Although our approach may have missed to classify about a quarter of surviving startups correctly, the predictions are of high practical value. The high precision indicates that the startups that are classified as survivors by our model have a high probability of actually surviving. Thus, our predictions are also highly relevant for practice. When being benchmarked against a naïve “majority rule” classification, our approach outperforms that baseline at least by factor three in that condition (i.e., 70% accuracy of our model vs. 20% accuracy of a naïve classification). Figure 1 further reflects a variable importance plot. Variance importance can be interpreted as the relative contribution of a variable to the prediction; the higher this metric, the

higher a variable's importance. The variable importance values were highly consistent across all survival rates such that we depict average variable importance in Figure 2 only. Our analysis shows that the perception of a startup's product in social media, i.e., Twitter Likes, seems to be the most important variable for making accurate survival predictions. We found the founders' character traits of openness and conscientiousness as well as the duration of their prior entrepreneurial experience to be quite important predictors. Moreover, the business focus of the startup seems to be important for discriminating between dead and alive startups.

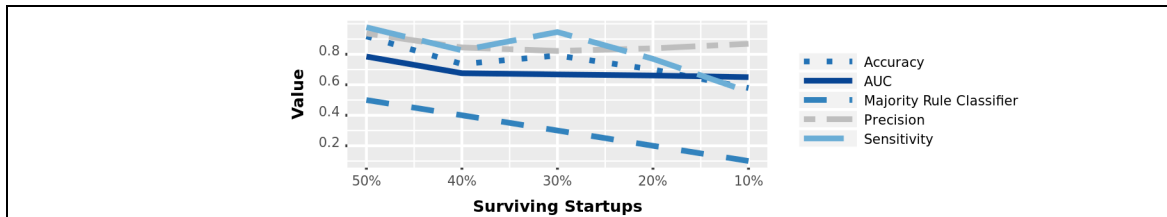


Figure 1. AUC, Accuracy, Recall and Precision of Survival Prediction

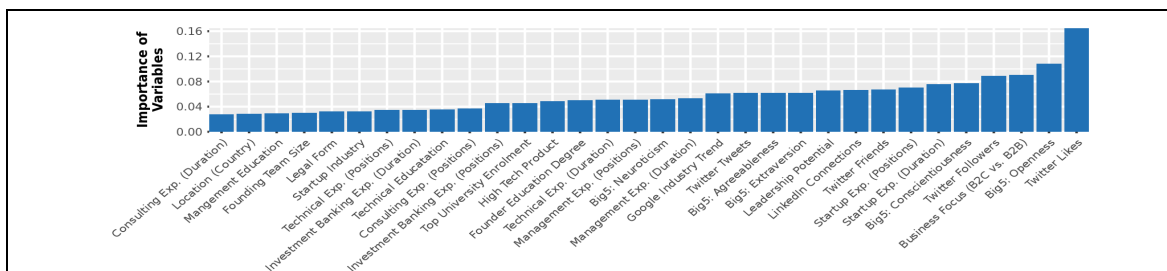


Figure 2. Relative Variable Importance of Survival Prediction

Next Steps and Expected Contributions

We will elaborate on our paper in three main areas. First, we will expand our database to other sources of digital footprints. We aim to collect data on financing rounds, web traffic and location data. Further, we will mine the entire Twitter communication of the investigated new ventures and will exploit the social media sentiment in order to capture a more accurate public perception. In this vein, one of the main challenges of our study is the comparability of time-stamps for different features in our model. This means that, for instance, new ventures activities on social media are never time-consistent. More specifically, some startups might directly create their Twitter account after founding, while others might create their account with introduction of their product. In order to mitigate this issue, we will increase the depth of our database to capture such information. Based on these new features, we will specifically investigate the impact of initial social media activities in the first few months after creation of the accounts. Second, based on the increased data sample, we will refine the quality of our measures that we conceptually derived from the entrepreneurship literature. For instance, we could follow Johannessen et al. (2001) and build an individual innovation measure, rating different questions from an “innovation-as-newness”-perspective. In this regard, we will also put emphasis on elaborating on a small set of features that contribute most to make accurate predictions such that we can increase the value of our approach. Third, we will elaborate on our planned sensitivity analysis. Given the data on the actual investments of two BA investment platforms, we will create a benchmark against which to compare the results of our study. To create such as baseline, we will compare the performance of our predictive models with the investment performance of actual investors. In such a simulation, we create hypothetical portfolios based on actual investment cases and compare the returns of those portfolios to the ones that 339 BAs have built over the past eight years.

We expect our study to make three main contributions to existing research and practice. First, we conceptualize empirical findings from the startup survival literature and provide a taxonomy of digital traces that can be used to predict early stage startup survival. While our measures are not novel per se for the entrepreneurship literature – various existing studies have already used measures to capture traits of entrepreneurs, their startups and their environment to predict startup success – our study goes beyond the

established work by offering a complete reconceptualization of these measures as digital traces that can be found on the Internet. As existing research on startup survival has been primarily centered on survey-based research or investigating trade register data (e.g., Raz and Gloor 2007) our work has the potential to introduce an entirely new set of measures to the domain of entrepreneurship research. Second, we provide robust empirical insights into the potential value of digital traces for predicting startup survival and potentially also their performance. By benchmarking our approach against a baseline of real BA investments, we will be able to find out if and how mining and processing digital traces might improve the predictive accuracy of BAs. Further, these insights will help identify a manageable set of robust features which show highest predictive accuracy. In so doing, we will might be able to add a novel perspective on survival factors of startups that may help elucidating highly conflicting results in existing research (e.g., Gimeno et al. 1997). Third, our study will expand current IS research on the business value of collecting and combining publicly available, unstructured and messy digital traces that individuals leave on the Internet.

References

- Asur, S., and Huberman, B. A. 2010. "Predicting the Future with Social Media," *International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, pp. 492-499.
- Audretsch, D. B., and Mahmood, T. 1994. "The Rate of Hazard Confronting New Firms and Plants in Us Manufacturing," *Review of Industrial Organization* (9:1), pp. 41-56.
- Bayus, B. L., and Agarwal, R. 2007. "The Role of Pre-Entry Experience, Entry Timing, and Product Technology Strategies in Explaining Firm Survival," *Management Science* (53:12), pp. 1887-1902.
- Bickel, P. J., and Freedman, D. A. 1984. "Asymptotic Normality and the Bootstrap in Stratified Sampling," *Annals of Statistics* (12:2), pp. 470-482.
- Blattberg, R. C., and Hoch, S. J. 1990. "Database Models and Managerial Intuition: 50% Model+ 50% Manager," *Management Science* (36:8), pp. 887-899.
- Böhm, M., Weking, J., Fortunat, F., Müller, S., Welpel, I., and Krcmar, H. 2017. "The Business Model DNA: Towards an Approach for Predicting Business Model Success,".
- Brüderl, J., Preisendörfer, P., and Ziegler, R. 1992. "Survival Chances of Newly Founded Business Organizations," *American Sociological Review* (57:2), pp. 227-242.
- Choi, H., and Varian, H. 2012. "Predicting the Present with Google Trends," *Economic Record* (88:1), pp. 2-9.
- Ciavarella, M. A., Buchholtz, A. K., Riordan, C. M., Gatewood, R. D., and Stokes, G. S. 2004. "The Big Five and Venture Survival: Is There a Linkage?," *Journal of Business Venturing* (19:4), pp. 465-483.
- Colombelli, A., Krafft, J., and Quatraro, F. 2013. "Properties of Knowledge Base and Firm Survival: Evidence from a Sample of French Manufacturing Firms," *Tech. Forecasting and Social Change* (80:8), pp. 1469-1483.
- Creamer, G. G., Ren, Y., Sakamoto, Y., and Nickerson, J. V. 2016. "A Textual Analysis Algorithm for the Equity Market: The European Case," *Journal of Investing* (25:3), pp. 105-116.
- Dellermann, D., Lipusch, N., Ebel, P. A., Popp, K. M., and Leimeister, J. M. 2017. "Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method," *International Conference on Information Systems (ICIS)*. Seoul.
- Delmar, F., and Shane, S. 2006. "Does Experience Matter? The Effect of Founding Team Experience on the Survival and Sales of Newly Founded Ventures," *Strategic Organization* (4:3), pp. 215-247.
- Fast, L. A., & Funder, D. C. 2008. "Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior," *Journal of Personality and Social Psychology*, 94(2), pp. 334-346.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* (17:3), pp. 37-53.
- Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* (29:5), pp. 1189-1232.
- Gloor, P. A., Dorsaz, P., Fuehres, H., and Vogel, M. 2013. "Choosing the Right Friends—Predicting Success of Startup Entrepreneurs and Innovators through Their Online Social Network Structure," *International Journal of Organisational Design and Engineering* (3:1), pp. 67-85.
- Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. 2005. "The Predictive Power of Online Chatter," *International Conference on Knowledge Discovery in Data Mining*, Chicago, pp. 78-87.

- Headd, B. 2003. "Redefining Business Success: Distinguishing between Closure and Failure," *Small Business Economics* (21:1), pp. 51-61.
- Hochberg, Y. V., Ljungqvist, A., and Lu, Y. 2007. "Whom You Know Matters: Venture Capital Networks and Investment Performance," *Journal of Finance* (62:1), pp. 251-301.
- Hunter, D. S., and Zaman, T. 2017. "Picking Winners: A Framework for Venture Capital Investment," *arXiv:1706.04229*.
- Hyytinen, A., Pajarinen, M., and Rouvinen, P. 2015. "Does Innovativeness Reduce Startup Survival Rates?," *Journal of Business Venturing* (30:4), pp. 564-581.
- Johannessen, J.-A.; Olsen, B., and Lumpkin, G. T. 2001. "Innovation as newness: what is new, how new, and new to whom?," *European Journal of Innovation Management* (4:1), pp. 20-31.
- Kane, T. J. 2010. "The Importance of Startups in Job Creation and Job Destruction,".
- Ko, E.-J., and McKelvie, A. 2018. "Signaling for More Money: The Roles of Founders' Human Capital and Investor Prominence in Resource Acquisition across Different Stages of Firm Development," *Journal of Business Venturing* (33:4), pp. 438-454.
- Kosinski, M., Stillwell, D., and Graepel, T. 2013. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences* (110:15), pp. 5802-5805.
- Krishna, A., Agrawal, A., and Choudhary, A. 2016. "Predicting the Outcome of Startups: Less Failure, More Success," *International Conference on Data Mining Workshops*, Barcelona, pp. 798-805.
- Lambiotte, R., and Kosinski, M. 2014. "Tracking the Digital Footprints of Personality," *Proceedings of the IEEE* (102:12), pp. 1934-1939.
- Marcus, B., Machilek, F., and Schütz, A. 2006. "Personality in Cyberspace: Personal Web Sites as Media for Personality Expressions and Impressions," *Journal of Personality and Social Psychology* (90:6), pp. 1014-1031.
- Mayer-Haug, K., Read, S., Brinckmann, J., Dew, N., and Grichnik, D. 2013. "Entrepreneurial Talent and Venture Performance: A Meta-Analytic Investigation of Smes," *Research Policy* (42:6), pp. 1251-1273.
- Mollick, E. 2016. "Containing Multitudes: The Many Impacts of Kickstarter Funding,". Available at SSRN: <https://ssrn.com/abstract=2808000>.
- Mollick, E. and Kuppuswamy, V. 2014. "After the Campaign: Outcomes of Crowdfunding," *UNC Kenan-Flagler Research Paper No. 2376997*. Available at SSRN: <https://ssrn.com/abstract=2376997>.
- Morrisette, S. G. 2007. "A Profile of Angel Investors," *Journal of Private Equity* (10:3), pp. 52-66.
- Mudambi, R., and Zahra, S. A. 2007. "The Survival of International New Ventures," *Journal of International Business Studies* (38:2), pp. 333-352.
- Muller, D., Te, F., Meyer, F., and Cvijikj, I. P. 2016. "Towards Data Driven Decision Support for Financial Institutions: Predicting Small Companies Business Volume in Switzerland," *International Conference on Computer Science and Information Technology*, Amman, pp. 1-6.
- Pe'er, A., and Keil, T. 2013. "Are All Startups Affected Similarly by Clusters? Agglomeration, Competition, Firm Heterogeneity, and Survival," *Journal of Business Venturing* (28:3), pp. 354-372.
- Petersen, M. A. 2004. *Information: Hard and Soft*. New York: Mimeo.
- Prowse, S. 1998. "Angel Investors and the Market for Angel Investments," *Journal of Banking & Finance* (22:6-8), pp. 785-792.
- Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J. 2012. "The Personality of Popular Facebook Users," *Conference on Computer Supported Cooperative Work*, Seattle, pp. 955-964.
- Raz, O., and Gloor, P. A. 2007. "Size Really Matters—New Insights for Start-Ups' Survival," *Management Science* (53:2), pp. 169-177.
- Riedl, C., Blohm, I., Leimeister, J. M., Krcmar, H. 2013. "The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities," *International Journal of Electronic Commerce* 17(3), pp. 7-36.
- Song, M., Podoyntsyna, K., Van Der Bij, H., and Halman, J. I. 2008. "Success Factors in New Ventures: A Meta-Analysis," *Journal of Product Innovation Management* (25:1), pp. 7-27.
- Wennberg, K., Wiklund, J., DeTienne, D. R., and Cardon, M. S. 2010. "Reconceptualizing Entrepreneurial Exit: Divergent Exit Routes and Their Drivers," *Journal of Business Venturing* (25:4), pp. 361-375.
- Yang, S., and Berger, R. 2017. "Relation between Start-Ups' Online Social Media Presence and Fundraising," *Journal of Science and Technology Policy Management* (8:2), pp. 161-180.