Journal of Business Venturing Insights 11 (2018) e00109

Contents lists available at ScienceDirect

Journal of Business Venturing Insights

journal homepage: www.elsevier.com/locate/jbvi

Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy

Torben Antretter[a], Ivo Blohm[b], Dietmar Grichnik[a], Joakim Wincent[a,c,d,*]

[a] University of St.Gallen, Dufourstrasse 40a, St. Gallen CH-9000, Switzerland
[b] University of St.Gallen, Müller-Friedberg-Strasse 8, St. Gallen CH-9000, Switzerland
[c] Hanken School of Economics, Arkadiankatu 22, Helsinki FI-00101, Finland
[d] Luleå University of Technology, Luleå 971 87, Sweden

ABSTRACT

Research indicates that interactions on social media can reveal remarkably valid predictions about future events. In this study, we show that online legitimacy as a measure of social appreciation based on Twitter content can be used to accurately predict new venture survival. Specifically, we analyze more than 187,000 tweets from 253 new ventures' Twitter accounts using context-specific machine learning approaches. Our findings suggest that we can correctly discriminate failed ventures from surviving ventures in up to 76% of cases. With this study, we contribute to the ongoing discussion on the importance of building legitimacy online and provide an account of how to use machine learning methodologies in entrepreneurship research.

1. Introduction

Previous studies have suggested that up to 78% of new ventures fail within the first five years of their existence (e.g., Song et al., 2008). Given these high failure rates, predicting survival has attracted attention in the entrepreneurship and management literature (e.g., Cooper, 1993; Gartner et al., 1999; Hyytinen et al., 2015). However, predicting new venture survival is challenging because of the outcome depends heavily on environmental developments and the specific complexity of each venture (Cooper, 1993). A widely used concept in understanding new ventures' chance of survival is their ability to build legitimacy. Legitimacy can be generally understood as "social judgment of acceptance, appropriateness, and desirability [that] enables organizations to access other resources needed to survive and grow" (Zimmerman and Zeitz, 2002: 414). In recent years, the process of building legitimacy has moved to a certain extent from the offline to the online world (e.g., Castelló et al., 2016; Etter et al., 2018). Although some studies on the relationship between legitimacy and survival have recognized the role of virtual embeddedness and online internet acceptance (e.g., Morse et al., 2007), the vast majority of research has explored and measured legitimacy in offline settings (e.g., Zimmerman and Zeitz, 2002). Thus, for the purposes of this paper, we use the term online legitimacy in referring to a subconcept of legitimacy, namely the social appreciation and/or desirability that is build and can be measured in the online world. The lack of online legitimacy research is surprising given that online social media has significantly changed the communication of and interaction between new ventures and their stakeholders (e.g., Kadam and Ayarekar, 2014; Yang and Berger, 2017).

* Corresponding author.
E-mail addresses: torben.antretter@unisg.ch (T. Antretter), ivo.blohm@unisg.ch (I. Blohm), dietmar.grichnik@unisg.ch (D. Grichnik), joakim.wincent@hanken.fi, joakim.vincent@unisg.ch (J. Wincent).

http://dx.doi.org/10.1016/j.jbvi.2018.e00109
Received 25 September 2018; Received in revised form 4 December 2018; Accepted 17 December 2018

Twitter,[1] in particular, has gained increasing popularity in entrepreneurship research (e.g., Kuppuswamy and Bayus, 2017) and a variety of other disciplines not only for explaining current events but also for predicting future outcomes using crowd judgments (for an overview: Schoen et al., 2013). For example, Asur and Huberman (2010) used movie reviews on Twitter to accurately predict movie sales in the first weeks after the respective movies were released. Given these insights, it is likely that a new venture's behavior on Twitter and the market's response to such behavior (i.e., measured through likes, followers, and the sentiment of user comments) provide a valid measure of online legitimacy and may thus be used to predict new venture survival. Scholarly attempts to leverage Twitter on a large scale to predict entrepreneurial outcomes and eventually new venture survival, however, have been limited (e.g., Obschonka et al., 2017).

We address this gap in the literature by building predictive models using data from new ventures' Twitter accounts. Furthermore, we provide an account of how to use data mining, natural language processing, and machine learning techniques in entrepreneurship research. Our results using random forest and gradient boosting classification models suggest that we can predict five-year survival with an accuracy of up to 76%. More specifically, we find that the average length of tweets, number of likes given, and number of likes received are among the most important predictors of new venture survival. As such, in addition to presenting an online measurement for legitimacy, we believe this study builds a strong case to support venture capital (VC) investment decisions as VC portfolios regularly show failure rates of up to 44% (Manigart et al., 2002). Our predictive model could, ceteris paribus, lead VC investors to decrease the failure rates in their portfolios (see Fig. 1 for a comparison). Our study makes two additional contributions. First, it adds to the emerging discussion about the digital web and its influence on new venture legitimacy (Castelló et al., 2016; Colleoni, 2013; e.g., Etter et al., 2018; Morse et al., 2007) by building and testing Twitter-based measures of online legitimacy, which could be used for future studies. Second, it contributes to the new venture survival literature (e.g., Boyer and Blazy, 2014; Hyytinen et al., 2015; Stenholm and Renko, 2016) by highlighting a new perspective on how to approach survival predictions based on large data samples generated via data and text mining and by discussing how such techniques can be combined with different machine learning methods to build predictive models. Besides other studies using online data to predict entrepreneurs' personality characteristics (e.g., Obschonka et al., 2017) this is – to the best of our knowledge – the first study in the entrepreneurship field showing the potential of using data mining, natural language processing, and machine learning to capture online information to predict entrepreneurial outcomes, such as survival.

2. Conceptual framework

2.1. The link between social media and legitimacy

The rise of social media has changed the way we need to look at new venture legitimacy in at least two ways. First, it enables new ventures to share information on a global scale and at minimum costs, thus allowing them to increase the effectiveness of their marketing communication (e.g., Kozinets et al., 2010). Studies have shown that social media campaigns can create social contagion and word-of-mouth "buzz" that drive product adoption and sales (Aral and Walker, 2011). Moreover, social media increases new ventures' ability to create and maintain weak ties, allowing them to manage a relatively large number of connections at the same time (Yang and Berger, 2017). Given these developments, scholars have recently started to grasp the role of social media in different steps of the venture-creation and legitimacy-building process. For instance, Yang and Berger (2017) discussed the use of social media for fundraising purposes and suggested that there is a relationship between popularity on Twitter and Facebook and the amount of funding new ventures receive. Moreover, a growing number of investors are considering online legitimacy in their funding decisions (e.g., by checking how many followers a company has and how creative it is in managing its online presence). An impressive social media effort may thus demonstrate legitimacy by showing that a firm has unique brand value and knows how to attract a fan base of customers and other stakeholders (Hong, 2013; Liang and Yuan, 2016).

Second, social media allows the public to provide direct and unfiltered feedback toward new ventures. While traditional newspapers and opinion pages only offer limited possibilities to express personal judgments (Lee and Carroll, 2011), social media, such as Twitter, has created public arenas where organizational activities are continuously discussed and evaluated (Whelan et al., 2013). These evaluations can have significant impacts on firm performance (George et al., 2014). A mere tweet from a trusted source can cause chain reactions in the press, social networks, and the stock market. In a recent study, Etter et al. (2018) discussed the potential of sentiment analysis to capture citizens' judgments on Twitter and argued that a legitimacy measure based on social media could "contribute to a more encompassing understanding of legitimacy" (2018:62). However, quantitative approaches using online legitimacy to research entrepreneurial outcomes are scarce. We therefore propose a way how to measure online legitimacy and test its value for predicting new venture survival.

2.2. Measuring online legitimacy in social media

Given the insights provided by prior literature, a measure of online legitimacy should include the following aspects. First, the amount of information that a firm provides about its products, management, and organization is considered a primary source of

[1] Twitter is a micro-blogging service founded in 2006. It allows users to post short messages (tweets) up to 280 characters in length and reply to and/or forward (re-tweet) other users' posts. Besides plain text, tweets may include hyperlinks to other websites, blogs, or pictures. Most tweets are publicly available. By December 2017, the average number of monthly active users worldwide was 330 million.

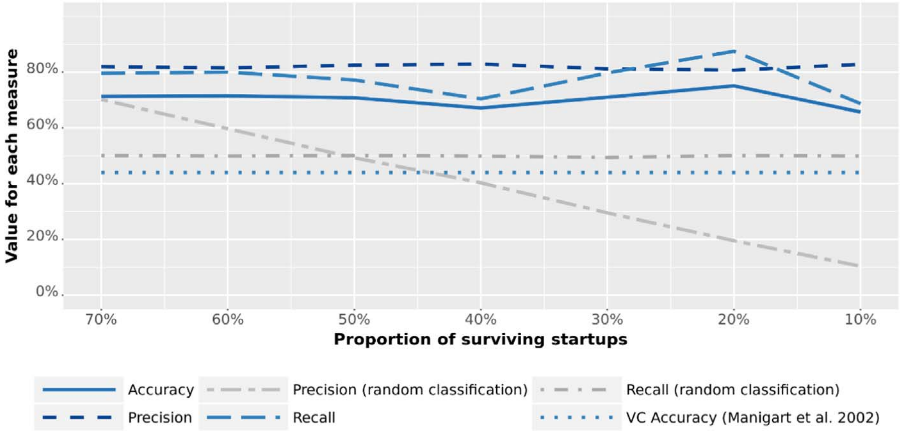

Fig. 1. Accuracy, sensitivity, and specificity of survival prediction.

legitimacy because potential customers often lack a frame of reference for understanding the benefits of a new venture's product (Shepherd and Zacharakis, 2003). Building on this argument, marketing scholars have pointed out that the frequency of firms' social media usage and the amount of information they provide through such channels are positively associated with product sales (e.g., Clark and Melancon, 2013; Reuber and Fischer, 2011). Assuming that new ventures use social media as a tool to create online legitimacy, the effort that they show to maintain relationships and provide information to their stakeholders should thus reflect their legitimacy (Kuppuswamy and Bayus, 2017).

Second, research has argued that the content of such posts is a valid proxy for new ventures' online reputation (Reuber and Fischer, 2011). As reputation – the relative position of an organization amongst its counterparts – is closely linked to legitimacy – which is often described as the individual perception of an entity – (e.g., Deephouse and Carter, 2005), we expect that not only the quantity, but also the content of and emotions with which information are shared on Twitter impacts online legitimacy. Sharing emotional content via Twitter may thus be used to create transparency and trustworthiness in order to gain, maintain and defend legitimacy (Ashforth and Gibbs, 1990). For instance, studies have shown that the emotional valence language on Twitter is highly correlated to the IPO performance of high-growth ventures (Liew and Wang, 2016).

Finally, in addition to new ventures' behavior on Twitter, the public's response to such behavior reflects stakeholders' unfiltered voices and opinions, thus indicating online legitimacy (Etter et al., 2018). Previous studies have shown that the degree of user interaction and the number of Twitter likes are valid proxies for the quality of firms' relationships with their customers (e.g., Clark and Melancon, 2013; Kadam and Ayarekar, 2014). Thus, the number of active fans or followers who frequently share, like, or comment on a new venture's online content likely contributes to a valid measure of the venture's online legitimacy.

3. Data, variables, and methods

3.1. Data: new ventures' social media profiles

To empirically assess the predictive power of online legitimacy, we obtained data on new ventures from a large early-stage investment platform in Switzerland. The companies were all seed or early-stage companies founded between 2006 and 2018. The main database included the names, websites, and social media profiles of each company. Our final sample included 253 firms for which we could infer five-year survival and that had a Twitter account. In line with prior research samples, 72% of the ventures in our sample survived for at least five years (Mudambi and Zahra, 2007: 71%). Following Kuppuswamy and Bayus (2017), we used Twitter's REST API to collect the ventures' tweets and related activity from the time they registered on Twitter to June 2018. This process resulted in a total of 187,323 tweets, 102,501 retweets, and 441,583 likes.

3.2. Variables

As outlined in the previous section, we focus on three different factors of online legitimacy: (1) quantity of information, (2) information content, and (3) interaction and confirmation metrics.

The outcome variable in our study is five-year survival. Thus, we measured whether a company actively offered a product or service for the time of five years after its incorporation. Since obtaining accurate data about new ventures is frequently hindered by uneven record keeping, lack of historical information, and potential source biases (Brush and Vanderwerf, 1992), scholars have described survival as the most reliable performance measure to study early-stage ventures (e.g., Gimeno et al., 1997). We measured survival as a dummy variable indicating whether a company was active for at least five years (1) or not (0). To operationalized new venture survival, we followed Raz and Gloor (2007) and performed an automated company search on the Internet to determine whether each startup still had a functioning website and, if so, what the company's website said. More specifically, when the company's website returned an error, was blank or displayed a notice, which indicated that the new venture was out of business, we inferred that the company was no longer actively providing products or services in the marketplace.

To identify positive and negative judgments that may contribute to our understanding of online legitimacy, we applied multilingual dictionary-based sentiment analysis, which is increasingly being used in management and entrepreneurship research (e.g., Nadkarni and Chen, 2014; Obschonka et al., 2017). We used the Linguistic Inquiry Word Count (LIWC) to capture positive and negative emotions in English, French, German, Spanish, Italian, Portuguese and Dutch Twitter messages (Pennebaker et al., 2001; Tausczik and Pennebaker, 2010).[2] As such, we were able to analyze more than 96% of the startups' entire Twitter communication. Based on this data, we also created a Naïve Bayes classifier that calculates the probability of a new venture surviving five years given the words used (and not used) in their Twitter communication. That is, for each of these seven languages, we identified the words that are most strongly associated with survival or death (Hotho et al., 2005). Furthermore, we created language dummies that isolate the effect of using different languages in a new ventures Twitter communication. An overview of all Twitter variables used in our model, their measurements, and their descriptive statistics are shown in Table 1. All values above 1000 were rounded to the nearest integer.

3.3. *Method: random forest and gradient boosting*

We used random forest and gradient boosting to build and train our predictive models. Both approaches are based on classification and regression trees (CART) (Breiman et al., 1984; Parisot et al., 2015). Classification trees determine the membership of an entity (e.g., a new venture) in given classes (e.g., dead or alive). Such models are represented by a directed graph composed of nodes, leaves, and branches. Each node represents a variable that is used for making predictions, and each node is followed by a branch that specifies a test on the value of this variable (e.g., if the number of followers surpasses a given value, a startup is classified as alive, and dead otherwise). The branches then result in new nodes or leaves that reflect a final classification (Parisot et al., 2015). Consequently, such trees can be considered as a set of "decision rules" for classifying new ventures as dead or alive.

As individual classification trees are prone to overfitting, the random forest algorithm creates a large number of randomly created classification trees. The idea is to randomly resample the data repeatedly to train a new classifier for each subsample with a random subsample of available variables. As different classifiers overfit the data in dissimilar ways, they are averaged out on a large scale (Liaw and Wiener, 2002). Besides its robustness against overfitting, the random forest algorithm is very user friendly as it requires the researcher to determine two main parameters (i.e., the number of variables used for building the individual trees and the number of trees) and is usually not very sensitive to their values. The idea behind gradient boosting (Friedman, 2001) is to add one classifier at a time so that each classifier is trained to improve the already existing ensemble of classification trees. In so doing, the gradient-boosted tree algorithm tries to find optimal combinations of trees in relation to a given set of training data in an attempt to learn from past prediction errors. Notice that for random forest, each classifier is trained independently from the rest. The main advantages of gradient boosting are that it has high predictive accuracy in many contexts and handles computational resources efficiently.

To simulate the potential of *online legitimacy* to predict new venture survival and to mitigate potential sampling biases from our data-collection procedure, we applied stratified bootstrapping (Bickel and Freedman, 1984). Bootstrapping relies on the logic of resampling from the original dataset to approximate the actual distribution of parameters (Efron and Tibshirani, 1994). In accordance with extant statistics, we defined survival rates for our samples ranging from 10% to 70%. For each survival rate, we created 2000 bootstrap samples. In each sample, we randomly drew 150, 200, and 250 new ventures with replacement from our data set of 253 new ventures. We stratified individual bootstrap samples according to the different survival rates. For example, when drawing 250 new ventures with a 70% survival rate, we picked 175 companies that survived for five years and 75 companies that did not survive. 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 the cases of the training set whose composition of dead and survived new ventures was constructed according to the defined survival rate. In order to measure the performance of our approach, we calculated several measures on the testing data sets. For investigating overall model performance, we first calculated the accuracy of our predictions (the percentage of startups that were correctly classified as "dead" or "alive"). However, due to the "accuracy paradox"[3] we also calculated the recall (the percentage of survived startups that have been correctly classified as "alive") and precision (the percentage of startups that were classified as "alive" and that have actually survived). 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. As all results are comparable across different analyses, we report the results for gradient boosting only with a bootstrap sample size of 200 new ventures. All computations were made in R using the "xgboost" package.

[2] LIWC also offers other libraries for other highly relevant variables for our study. Unfortunately, many of them are only available in the English language. Given the multilinguistic scope of our study, we only use the LIWC standard library that has been translated to various languages. In greater detail, we used only the affect words.

[3] Accuracy may reflect a misleading measure of a model's actual performance. Consider the case that a model classifies all startups as "dead". If we assume a survival rate of 10% and use this model for classifying 100 startups, the model's performance would result in an accuracy of 90% (90 dead startups that have been classified correctly / 100 classified startups). While the accuracy of this model seems to be high, the model itself is useless, because it cannot identify any surviving company.

Table 1
Description of variables.

Variable	Measurement	Mean	SD	Min.	Max.
Quantity of information					
Provides description	Dummy (=1) if user-defined company description contains at least one character	0.86	0.35	0	1
Provides weblink	Dummy (=1) if user-defined weblink contains at least one character	0.87	0.34	0	1
Provides location	Dummy (=1) if user-defined location contains at least one character	0.82	0.38	0	1
Length of tweets	Avg. length of characters for each of the new ventures' tweets	102.06	31.29	0	174.18
No. of tweets	Total no. of tweets since the incorporation of the account	1120	3031	0	38,341
Tweet languages	Total no. of languages used in Tweets	6.83	5.26	0	27
Tweets in English	Dummy indicating tweets in English	0.83	0.37	0	1
Tweets in German	Dummy indicating tweets in German	0.47	0.50	0	1
Tweets in French	Dummy indicating tweets in French	0.35	0.48	0	1
Tweets in Italian	Dummy indicating tweets in Italian	0.07	0.26	0	1
Tweets in Spanish	Dummy indicating tweets in Spanish	0.15	0.35	0	1
Tweets in Portuguese	Dummy indicating tweets in Portuguese	0.15	0.35	0	1
Tweets in Dutch	Dummy indicating tweets in Dutch	0.03	0.18	0	1
Content analysis					
Sharing pos. content tweets	Avg. emotional valence of tweets for all languages	3.65	2.57	0	20
Sharing neg. content tweets	Avg. emotional valence of tweets for all languages	0.42	0.4	0	2.59
Sharing pos. content retweets	Avg. emotional valence of retweets for all languages	2.9	1.89	0	10.53
Sharing neg. content retweets	Avg. emotional valence of retweets for all languages	0.39	0.42	0	2.8
Sharing pos. content replies	Avg. emotional valence of replies for all languages	6.48	5.58	0	33.33
Sharing neg. content replies	Avg. emotional valence of replies for all languages	0.49	1.09	0	14.29
Words used	Avg. survival probability based on Naïve Bayes classification for all languages	0.6	0.42	0	1
Interaction/confirmation metrics					
No. of likes received	No. of likes received since incorporation	512.05	1227	0	11,924
No. of retweets given	No. of retweets performed since incorporation	123.88	239.46	0	1522
No. of retweets received	No. of retweets by others since incorporation	281.26	512.97	0	3114
No. of followings	No. of accounts the new venture follows	617.08	1470	0	16,708
No. of listings	No. of user lists the account was added to since incorporation	49.85	141.39	0	1828
No. of replies	No. of replies given since incorporation	89.36	243.86	0	2702
No. of likes given	No. of likes given to others since incorporation	1233	6332	0	80,154
No. of followers	No. of followers accumulated since incorporation	1749	9063	0	102,252
No. of new followers	Average no. of new followers per day	0.73	3.21	0	34.43
User engagement	Sum of replies and retweets divided by no. of followers	51.7	306.75	0	3986
Followers per following	No. of followers divided by no. of followings	18.15	182.3	0	2764
Followers per tweet	No. of followers divided by total no. of tweets	3.06	13.64	0	193.66
No. of retweeters	No. of Twitter users that retweeted the venture's tweets since incorporation	41.56	54.1	0	272
No. of retweeters' followers	No. of followers of users that retweeted tweets[a]	171,650	320,385	0	2854,316
Followers per retweet	Avg. number of Twitter users that have been reached by a retweet	4807	9953	0	126,385

[a] We limited the mining of the number of retweeters' followers due to rate limits of Twitter's REST API. We only collected data on the retweeters' followers for the most frequently retweeted retweets.

4. Analysis and results

4.1. Survival prediction

Fig. 1 depicts the mean values of the performance measures across all bootstrap samples. Our results show that the classification models provide accurate results across all survival rates. The accuracy across all survival rates is about 74%, with a maximum of 76% for the 50% survival rate. However, high recall (mean 86%) and precision values (mean 80%) are more important for evaluating the value of our model. For instance, in the realistic case of 20% startup survival (Song et al., 2008), our model shows a recall of 81% and precision of 83%. Although our approach may have missed to classify about a fifth of surviving startups correctly, the predictions are of high practical value. The high precision indicates that the startups that are classified as survivors have a probability of 83% of actually surviving. When being benchmarked against a simulated random classification, our approach significantly outperforms that baseline. For the survival rate of 20%, random classification results in a recall of 50% and a precision of 19%. This results in relative performance improvements of 62% (recall) and 337% (precision). While precision is relatively constant across all survival rates, recall slightly decreases with smaller rates of survival in our sample. As such, it becomes more difficult for the trained model to accurately pick survived companies from the test dataset when the relative amount of survived companies decreases. As illustrated in Fig. 1, we believe that our results present a strong case to use online legitimacy to predict new venture survival, especially when our prediction accuracy is compared to the survival rates of portfolio companies of professional VC firms mentioned in previous research.

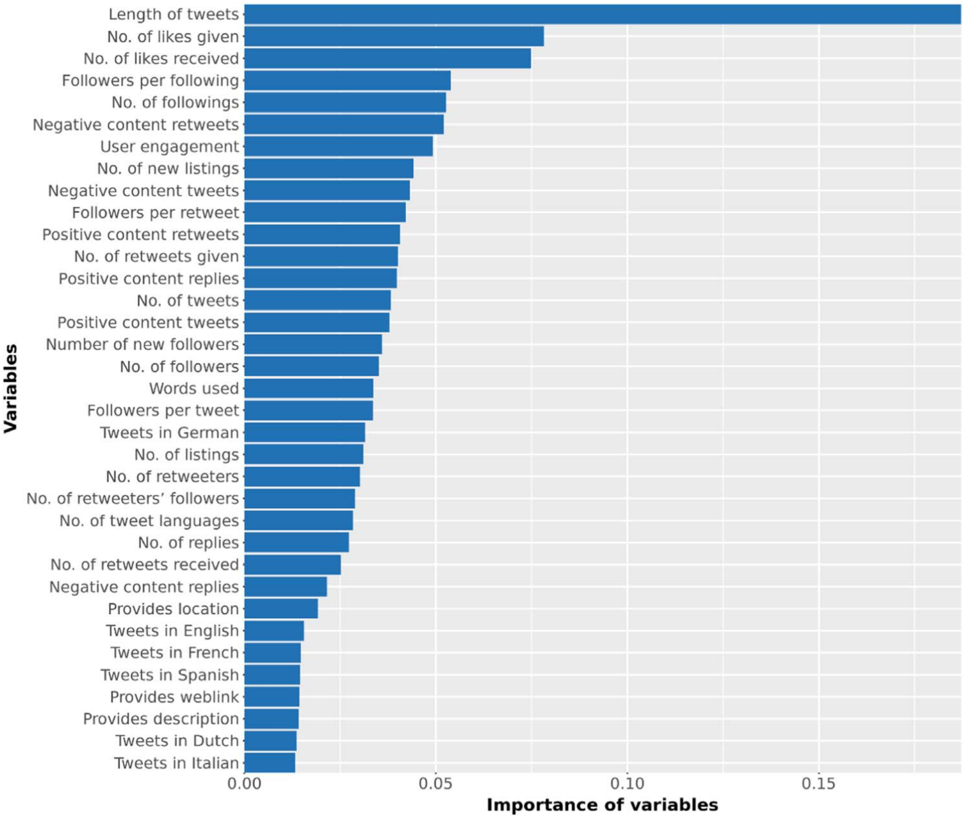

Fig. 2. Relative importance of legitimacy factors for our prediction.

4.2. Importance of individual online legitimacy factors for new venture survival

To get a more detailed understanding of the role of individual Twitter metrics in predicting new venture survival, we now look at the importance of each variable for the prediction model (see Fig. 2). “Variable importance” reflects the relative contribution of each measure to the overall prediction model. Thus, the higher this value, the higher the variable's importance. Our results show that the average length of tweets, which is limited to 280 characters, is the most important survival predictor in our model (Rank 1). This result suggests that the higher the quantity of information a new venture provides, the more online legitimacy it is likely to create. Interestingly, this logic only seems to apply to direct interactions with other users as the amount of information a new venture passively provides through its Twitter profile description (Rank 33) or a link to its company website (Rank 32) are among the weakest predictors in our model. The same is true for the dummies which reflect the languages in which a startup has tweeted. Regarding the content variables, we find that sharing negative content weights much higher (e.g., Ranks 6 and 9) than sharing positive content (Rank 11 and 15) in determining whether a company will fail or survive. Finally, our engagement variables show that it is not only other users' interactions with a new venture's content (i.e., number of likes received [Rank 3] or number of followers [Ranks 17]) that determine online legitimacy but also a new venture's interactions with others (i.e., number of likes given [Rank 2] or number of followings [Rank 5]).

5. Discussion and conclusion

In this paper, we investigate whether Twitter can be used to measure online legitimacy to predict new venture survival, one of the key performance measures for early-stage ventures. Our results show that using Twitter, we can predict five-year survival with an accuracy of up to 76%. More importantly, our approach shows high values of recall and precision such that the predictions have high practical value. Interestingly, some variables associated with the process of building online legitimacy (i.e., length of tweets and number of likes given) have higher predictive power in our model than the variables associated with social appreciation about a new venture (i.e., number of likes received, number of followers, etc.). This result is somewhat surprising as we would expect the publics' judgment to be a robust indicator of future performance (Kadam and Ayarekar, 2014).

However, studies on electronic word-of-mouth on social media have shown that user engagement strongly depends on network effects (e.g., Luarn et al., 2014). Thus, when it comes to seed and early-stage ventures, people might be less willing to display their public opinion on matters that have little relevance to their desire for social interaction (Hennig-Thurau et al., 2004). The predictive power of social judgments as a co-constructive part of online legitimacy may thus be much higher for later-stage ventures that have already accumulated a particular audience or fan base on Twitter.

In sum, our study makes four contributions to both research and practice. First, new ventures may use these insights to shape their strategic decisions on how to build legitimacy in a digital environment. Entrepreneurs can deliberately choose the amount and type of content they share via social media to increase their chances of survival. Second, VC investors – whose portfolios still show failure rates of up to 44% (Manigart et al., 2002) – can use the results of this study to decrease the downside risk in their portfolios and thus make more successful investment decisions. Third, our study contributes to the recent discussion of the role of social media in venturing (e.g., Gloor et al., 2013; Reuber and Fischer, 2011; Yang and Berger, 2017) by proposing an integrated measure of online legitimacy and testing its validity for new venture survival prediction. Finally, we add to the new venture survival literature (e.g., Boyer and Blazy, 2014; Hyytinen et al., 2015; Stenholm and Renko, 2016) by highlighting the potential of large data samples generated via text mining and machine learning methodologies to predict survival.

However, our study is not without limitations. Future studies exploring additional methods and online channels that allow new ventures to communicate and receive direct feedback, such as LinkedIn, Facebook, or Instagram, are essential to assess the validity of our results. Further, this study is the first to use Twitter to build an integrated measure of online legitimacy and test its predictive power in entrepreneurship research. This novel approach opens up new avenues for future research to investigate whether online legitimacy can be used to predict other entrepreneurial outcomes (i.e., valuations or exit probabilities). Above that, our measure does not capture whether online legitimacy leads to new resources in the same way as does our common understanding of legitimacy. As such, future research may consider investigating the resource acquisition effects resulting from higher levels of online legitimacy. Finally, future research could broaden our measure of online legitimacy by adding further business-related online metrics, such as web traffic, search volume, or digital media outlets.

Conflict of interest

None.

References

Aral, S., Walker, D., 2011. Creating social contagion through viral product design: a randomized trial of peer influence in networks. Manag. Sci. 57 (9), 1623–1639.

Ashforth, B.E., Gibbs, B.W., 1990. The double-edge of organizational legitimation. Organ. Sci. 1 (2), 177–194.

Asur, S., Huberman, B.A., 2010. Predicting the future with social media. *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, 1(1): 492-499.

Bickel, P.J., Freedman, D.A., 1984. Asymptotic normality and the bootstrap in stratified sampling. Ann. Stat. 12 (2), 470–482.

Boyer, T., Blazy, R., 2014. Born to be alive? The survival of innovative and non-innovative French micro-start-ups. Small Bus. Econ. 42 (4), 669–683.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press, New York Chapman & Hall.

Brush, C.G., Vanderwerf, P.A., 1992. A comparison of methods and sources for obtaining estimates of new venture performance. J. Bus. Ventur. 7 (2), 157–170.

Castelló, I., Etter, M., Årup Nielsen, F., 2016. Strategies of legitimacy through social media: the networked strategy. J. Manag. Stud. 53 (3), 402–432.

Clark, M., Melancon, J., 2013. The influence of social media investment on relational outcomes: a relationship marketing perspective. Int. J. Mark. Stud. 5 (4), 132–142.

Colleoni, E., 2013. CSR communication strategies for organizational legitimacy in social media. Corp. Commun. 18 (2), 228–248.

Cooper, A.C., 1993. Challenges in predicting new firm performance. J. Bus. Ventur. 8 (3), 241–253.

Deephouse, D.L., Carter, S.M., 2005. An examination of differences between organizational legitimacy and organizational reputation. J. Manag. Stud. 42 (2), 329–360.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press Book, Boca Raton.

Etter, M., Colleoni, E., Illia, L., Meggiorin, K., D'Eugenio, A., 2018. Measuring organizational legitimacy in social media: Assessing citizens' judgments with sentiment analysis. Bus. Soc. 57 (1), 60–97.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Gartner, W., Starr, J., Bhat, S., 1999. Predicting new venture survival: an analysis of "anatomy of a start-up." cases from Inc. magazine. J. Bus. Ventur. 14 (2), 215–232.

George, G., Haas, M.R., Pentland, A., 2014. Big data and management. Acad. Manag. J. 57 (2), 321–326.

Gimeno, J., Folta, T.B., Cooper, A.C., Woo, C.Y., 1997. Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. Adm. Sci. Q. 42 (4), 750–783.

Gloor, P.A., Dorsaz, P., Fuehres, H., Vogel, M., 2013. Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure. Int. J. Organ. Des. Eng. 3 (1), 67–85.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G., Gremler, D.D., 2004. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet? J. Interact. Mark. 18 (1), 38–52.

Hong, N. 2013. If you look good on twitter, VCs may take notice, *The Wall Street Journal*: 1-3. New York.

Hotho, A., Nürnberger, A., Paaß, G., 2005. *A brief survey of text mining*. Paper presented at the Ldv Forum.

Hyytinen, A., Pajarinen, M., Rouvinen, P., 2015. Does innovativeness reduce startup survival rates? J. Bus. Ventur. 30 (4), 564–581.

Kadam, A., Ayarekar, S., 2014. Impact of social media on entrepreneurship and entrepreneurial performance: special reference to small and medium scale enterprises. SIES J. Manag. 10 (1), 3–11.

Kozinets, R.V., De Valck, K., Wojnicki, A.C., Wilner, S.J., 2010. Networked narratives: understanding word-of-mouth marketing in online communities. J. Mark. 74 (2), 71–89.

Kuppuswamy, V., Bayus, B.L., 2017. Does my contribution to your crowdfunding project matter? J. Bus. Ventur. 32 (1), 72–89.

Lee, S.Y., Carroll, C.E., 2011. The emergence, variation, and evolution of corporate social responsibility in the public sphere, 1980–2004: the exposure of firms to public debate. J. Bus. Ethics 104 (1), 115–131.

Liang, Y.E., Yuan, S.-T.D., 2016. Predicting investor funding behavior using crunchbase social network features. Internet Res. 26 (1), 74–100.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2 (3), 18–22.

Liew, J.K.-S., Wang, G.Z., 2016. Twitter sentiment and IPO performance: a cross-sectional examination. J. Portf. Manag. 42 (4), 129–135.

Luarn, P., Yang, J.-C., Chiu, Y.-P., 2014. The network effect on information dissemination on social network sites. Comput. Hum. Behav. 37 (1), 1–8.

Manigart, S., Baeyens, K., Van Hyfte, W., 2002. The survival of venture capital backed companies. Ventur. Capital. 4 (2), 103–124.

Morse, E.A., Fowler, S.W., Lawrence, T.B., 2007. The impact of virtual embeddedness on new venture survival: overcoming the liabilities of newness. Entrep. Theory Pract. 31 (2), 139–159.

Mudambi, R., Zahra, S.A., 2007. The survival of international new ventures. J. Int. Bus. Stud. 38 (2), 333–352.

Nadkarni, S., Chen, J., 2014. Bridging yesterday, today, and tomorrow: CEO temporal focus, environmental dynamism, and rate of new product introduction. Acad. Manag. J. 57 (6), 1810–1833.

Obschonka, M., Fisch, C., Boyd, R., 2017. Using digital footprints in entrepreneurship research: a twitter-based personality analysis of superstar entrepreneurs and managers. J. Bus. Ventur. Insights 8 (1), 13–23.

Parisot, O., Didry, Y., Tamisier, T., Otjacques, B., 2015. Helping predictive analytics interpretation using regression trees and clustering perturbation. J. Decis. Syst. 24 (1), 55–72.

Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic Inquiry and Word Count (LIWC): LIWC 2001. Erlbaum, Mahwah.

Raz, O., Gloor, P.A., 2007. Size really matters—new insights for start-ups' survival. Manag. Sci. 53 (2), 169–177.

Reuber, A.R., Fischer, E., 2011. International entrepreneurship in internet-enabled markets. J. Bus. Ventur. 26 (6), 660–679.

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. Int. J. Electronic Commerce 17 (3), 7–36.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., Gloor, P., 2013. The power of prediction with social media. Internet Res. 23 (5), 528–543.

Shepherd, D.A., Zacharakis, A., 2003. A new venture's cognitive legitimacy: an assessment by customers. J. Small Bus. Manag. 41 (2), 148.

Song, M., Podoynitsyna, K., Van Der Bij, H., Halman, J.I., 2008. Success factors in new ventures: a meta-analysis. J. Product. Innov. Manag. 25 (1), 7–27.

Stenholm, P., Renko, M., 2016. Passionate bricoleurs and new venture survival. J. Bus. Ventur. 31 (5), 595–611.

Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: liwc and computerized text analysis methods. J. Lang. Soc. Psychol. 29 (1), 24–54.

Whelan, G., Moon, J., Grant, B., 2013. Corporations and citizenship arenas in the age of social media. J. Bus. Ethics 118 (4), 777–790.

Yang, S., Berger, R., 2017. Relation between start-ups' online social media presence and fundraising. J. Sci. Technol. Policy Manag. 8 (2), 161–180.

Zimmerman, M.A., Zeitz, G.J., 2002. Beyond survival: achieving new venture growth by building legitimacy. Acad. Manag. Rev. 27 (3), 414–431.