creativity and innovation management

How good are ideas identified by an automatic idea detection system?

Journal:	*Creativity and Innovation Management*
Manuscript ID	1587-CIM-A-12-2016.R2
Manuscript Type:	Special Issue Article
Keywords:	Idea quality, Text mining, Machine learning, Natural language processing, Big data, Online communities, Support vector machines, Crowdsourcing

How good are ideas identified by an automatic idea detection system?

Abstract

Online communities can be an attractive source of ideas for product and process innovations. However, innovative user-contributed ideas may be few. From a perspective of harnessing "big data" for inbound open innovation, the detection of good ideas in online communities is problem of detecting rare events. Recent advances in text analytics and machine learning have made it possible to screen vast amounts of online information and automatically detect user-contributed ideas. However, it is still uncertain whether the ideas identified by such systems will also be regarded as sufficiently novel, feasible and valuable by firms who might decide to develop them further. A validation study is reported in which 200 posts were extracted from an online community using the automatic idea detection system by Christensen, Nørskov, Frederiksen and Scholderer (2017; DOI: 10.1111/caim.12202). Two company professionals evaluated the posts in terms of idea content and idea quality. The results suggest that the automatic idea detection system is sufficiently valid to be deployed for the harvesting and initial screening of ideas and that the profile of the identified ideas (in terms of novelty, feasibility and value) follows the same pattern identified in studies of user ideation in general.

Introduction

Big data has been predicted to revolutionise innovation and how firms will create value for themselves, their customers and society (e.g., see McAfee & Brynjolfsson, 2012). Artificial intelligence systems that leverage big data allow more and more tasks to be solved in an automatic manner. Whilst in the past, these were predominantly tasks of a mundane and repetitive nature, advances in text analytics and machine learning have also made it possible to solve more complex problems (Christensen, Nørskov, Frederiksen, & Scholderer, 2017).

A problem that continues to occupy scholars and practitioners of new product development is how to obtain and select ideas for new products (e.g., di Gangi, Wasko, & Hooker, 2010; van den Ende, Frederiksen, & Prencipe, 2015; Frederiksen & Knudsen, 2017). In the context of inbound open innovation, Ooms, Bell and Kok (2015), for example, argue that firms can enhance their receptivity—i.e., their capacity to absorb more diverse external knowledge from more varied sources—by engaging with social media. Whilst this can in theory expand a firm's boundaries for information absorption, the extent of engagement with social media is still constrained by available staff time. Such constraints can to some degree be overcome if companies develop or adopt systems that automate parts of the absorption process.

The aim of the research presented here is to show how the performance of automated systems in areas such as inbound open innovation can be evaluated. On one hand, the study should be seen as a feasibility study of whether automated detection of ideas for product- and process innovations is actually possible. On the other hand, it should also be regarded as a validation study that probes the "veracity" and "value" aspects of big data (Gandomi & Haider, 2015) in the context of a specific application case.

Literature review

Big data

Big data has received much attention in recent years, but it is not a new concept as such. Big data can be seen as a product of digitalisation, the "digital footprint" of an electronically mediated reality (Zwitter, 2014). Others stress instrumental aspects, regarding big data as a tool for generating insights (Boyd & Crawford, 2012; Nunan & Di Domenico, 2013). We prefer to see big data as a *resource* whereas *tools* (such as text analytics, machine learning and other artificial intelligence techniques) help create value from the resource. As an analogy, one might think of big data as the *oil* and of artificial intelligence as the *combustion engine* that makes the oil useful. In technical terms, big data refers to databases that are too big to be handled by conventional data warehousing systems. The "bigness" of big data is often characterised in terms of three parameters: variety, velocity and volume (Hsinchun Chen, Chiang, & Storey, 2012). Variety refers to the heterogeneity of data types: part of the database content may be structured and numeric (e.g., transaction data from retail channels) but other parts may have different forms, for example free text, image and video files exchanged on social media networks. Velocity refers to how fast new data is being generated. In order to utilise newly generated data, it be retrieved on a continuous basis. Volume refers to the amount of data, measured in terms rows and columns, complexity of databases, and total storage volume.

Online communities as idea reservoirs

One domain where big data has been generated since the early days of the Internet is online communities on message board systems and social media networks. Online communities where users exchange experiences and discuss potential improvements for new products and processes have been identified as rich reservoirs of ideas that can fuel the innovation processes of firms (Van de Ven, 1986; Ekvall, 1997; Vandenbosch, Saatcioglu, & Fay, 2006; van den Ende, Frederiksen, & Prencipe, 2015). Ideas do not have to originate from the creative mind of the firm's employees but can also originate from the users of its products, services and technologies (Kristensson, Gustafsson, & Archer, 2004; Magnusson, 2009; von Hippel, Ogawa, & de Jong, 2011; Poetz & Schreier, 2012; Majchrzak & Malhotra, 2013; Magnusson, Wästlund, & Netz, 2014).

Prominent examples of the role of user communities in open inbound innovation are the communities hosted by Dell (di Gangi, Wasko, & Hooker, 2010; Poetz & Schreier, 2012), Lego (Antorini, 2007; Antorini, Muñiz, & Askildsen, 2012; Nørskov, Antorini, & Jensen, 2015), Propellerhead (Jeppesen & Frederiksen, 2006) and IBM (Mahr & Lievens, 2012). Firm-hosted communities such as these have the advantage that the hosting firm can retain a certain degree of control. The communities are typically based on software that allows registered users to post ideas, comment on and vote for ideas posted by other users in a highly structured manner. The downside of this approach is that it requires an extensive base of committed product users or firm-loyal customers who have an intrinsic interest in suggesting ideas to the firm.

However, users do not only gather in firm-hosted communities. A vast amount of online communities exist that are firm-free (Füller, Bartl, Ernst, & Mühlbacher, 2006; Füller, Jawecki, & Mühlbacher, 2007). The most prominent cases include open-source software development communities such as those responsible for the Linux kernel, R and Python. These are examples of firm-free "products" and platforms that have developed in a distributed manner, utilising online collaboration tools such as GitHub and Sourceforge. The fact that the resulting products are now perfectly able to compete with their commercial counterparts (such as the products ranges of the SAS Institute or Microsoft) is a clear demonstration of the potential of such communities (von Krogh, Spaeth, & Lakhani, 2003; von Krogh & von Hippel, 2006)

The problem with firm-free communities is that they, unlike most firm-hosted communities, are usually *not* based on a crowdsourcing architecture that would enable easy harvesting and collaborative filtering of the community-generated ideas. Assigning employees to manual monitoring of community contributions is often the only viable solution if firms want to benefit from the ideas generated in firm-free communities. This is time-consuming and expensive; online communities may contain several hundred thousand posts and comments. The sheer amount of information in which the ideas are hidden is a practical barrier to finding the ideas and utilising them for innovation (Lin, Hsieh, & Chuang, 2009; Thorleuchter & Van den Poel, 2013).

Automatic idea detection

A new and efficient way of solving the needle-in-a-haystack problem is to use classification algorithms that can screen arbitrary amounts of community posts and comments and identify those that are likely to contain ideas. Using text analytics and machine learning methods, Christensen, Nørskov, Frederiksen, & Scholderer (2017) develop such an algorithm and demonstrate its classification performance and efficiency for the case of extracting new product ideas from an online community related to Lego. Christensen, Liland, et al. (2017) show that the same principles can be applied to extract ideas for innovations from a community related to craft brewing. The authors argue that their method is applicable across different technological areas and product categories because most people use a specific set of words and expressions when they communicate ideas to each other. That is, we humans have a very special discourse for talking about our ideas and problems. We humans recognize ideas, when we read and hear them, in the same manner as we recognize a car, when we see a car, and we recognize a ship when we see a ship. Since the presence of such linguistic markers can easily be detected in a given online community post or comment, it can also potentially be exploited in the screening of arbitrarily large collections of posts, comments or other types of semi- or unstructured text. If implemented as a screening tool in a company's R&D or marketing department, it can significantly reduce the labour costs that would arise if R&D staff were assigned to manual monitoring of community activity.

Aims of the study

We believe that the method introduced by Christensen et al. (2017) shows potential for aiding firms in their search for innovative ideas and may thus serve as a tool for extending the boundaries of inbound open innovation. Still, some questions must be addressed before such a method should be implemented in a firm's innovation processes. Ideas identified by the Christensen et al. (2017) method, for example, have not yet been evaluated by company-internal R&D or marketing staff and it remains to be investigated if ideas detected by such an automated system will also recognized as ideas by company staff. In addition, the ideas must be seen as sufficiently novel, feasible and valuable by the R&D or marketing staff who would be responsible to take the identified ideas further (e.g., development into concepts or prototypes). The aim of the present paper is to fill these two gaps. Specifically, we would like to contribute in two respects to the literature:

- Our first contribution is to assess whether ideas from an online community, identified by an artificial intelligence system such as the one described by Christensen et al. (2017), will also be perceived as ideas by company-internal staff.
- Our second contribution is to investigate if the ideas that are detected by the system will also be perceived as *good* ideas by company-internal staff.

There is a reason why believe these issues are important. We address potential acceptance problems that were also in the general innovation literature initially seen as barriers for the uptake of user-contributed ideas by companies. Since then, many studies have demonstrated that user-contributed ideas can often compete with the ideas generated by company-internal staff (see e.g. Kristensson et al., 2004; Magnusson, 2009; Magnusson et al., 2014) and therefore deserve to be given a fair chance. As a consequence, dedicated idea crowdsourcing systems have gained widespread acceptance in the business community. Poetz and Schreier (2012), for example, investigate if users ideas posted in a *firm-hosted, closed* idea-crowdsourcing community can compete with ideas generated by company professionals. Our study extends this question to the mode of automated idea-harvesting. The study we report asks if user-contributed ideas posted in an *open* online community, identified by an artificial intelligence-based system, can reach sufficient recognition among company professionals. The answer to this question will provide guidance for research and practice in the open innovation domain. If the answer is positive, artificial intelligence systems for idea-harvesting can be implemented in practice, and research can focus on optimising the methods. If the answer is negative, research can focus on unresolved problems and acceptance barriers related to automatic idea detection systems.

An online community related to craft brewing was used as the idea base for our study. A dataset consisting of 200 automatically extracted online community posts was generated for addressing the two aims. It is the first time results based on this particular dataset are reported in the literature. Employees of Norwegian craft brewery *Døgne Ø* evaluated the automatically extracted ideas.

Method

Machine learning for idea detection

The machine learning system we used for extracting the 200 texts is described in detail in Christensen, Nørskov et al. (2017) and Christensen, Liland, et al. (2017). Although the technical properties of the system are not the central focus of the present paper, we will give a brief description of the system and how it was employed in our study.

The machine learning system takes as input idea texts and non-idea texts that have been identified by human raters. The texts used for this study originate from *alt.beer.home-brewing*, a Usenet-based online community related to craft brewing. In this community people from all over the world discuss brewing-related issues. At the time the texts were extracted, the community contained altogether 10582 posts. 3000 of these were selected at random and extracted for the development of the training of the system (detailed results based on these 3000 texts have been reported in Christensen, Liland, et al. (2017)). Those that contained ideas were identified via crowdsourcing, using the *CrowdFlower* platform (a service similar to Amazon's *Mechanical Turk*). Five raters were assigned to each text and instructed to label the text as an idea text if it contained at least one idea.

Before the texts could be used for machine learning, several text pre-processing steps were performed. In this process, the raw text content was turned into a row-column format, where each text was represented as a row and each term (i.e., each unique word or expression) as a column. All numbers, punctuation marks and stop words were removed. Uni-grams, bi-grams and tri-grams were generated. All terms that did not occur in at least 0.2% of the texts were omitted from the analysis (this is a standard text cleaning step; e.g., see Antons, Kleer, & Salge (2016)). This process resulted in a dataset consisting of 10514 terms representing 10582 texts.

From the 3000 texts in the database, we excluded all texts where not all five CrowdFlower raters had agreed on the class membership. After excluding these, the new database contained 1393 texts. 405 of the texts were idea texts and 988 were non-idea texts. The texts were partitioned at random into three separate data sets: a training set (consisting of 70% of the texts), a validation set (15% of the texts) and a hold-out or test set (15% of the texts). Such a partition is essential for training a machine learning system (in the training set), the fine-tuning of its paramters (in the validation set) and for an unbiased evaluation of its performance on previously unseen data (hold-out). Based on the training set, validation set and hold-out, the automatic idea detection system was trained and tested. The system is based on a linear support vector machine classifier (for details, see Christensen, Liland, et al., 2017). Key performance statistics are reported in Table 1.

--- Table 1 ---

From the remaining 7582 texts (10582 – 3000 = 7582) which had not been involved in the training, validation and testing of the system, 200 texts were extracted for the present study by using the linear support vector machine classifier: 100 which the classifier had labelled as idea texts and 100 which the classifier had labelled as non-idea texts. A histogram of the posterior probability scores underlying these classifications is shown in Figure 1. These 200 texts were used in the present study as the idea and non-idea texts to be classified and rated by two brewing professionals.

--- Figure 1 ---

Measuring idea quality

The perceived quality of an idea can depend on the perspective of the person evaluating the idea. This topic has received much attention in the creativity and innovation management literature. In principle, idea quality can be measured on a “good idea” to “bad idea” scale, but in most research it is decomposed into several attributes that represent conceptually distinct dimensions of quality. Dean, Hender, Rodgers and Santanen (2006) provide a comprehensive review of the idea quality literature published between 1990 and 2005. Based on the altogether 90 identified studies, they suggest that four dimensions of idea quality can be distinguished: novelty, workability, relevance and specificity. An idea is novel if it contains something that is new. An idea is workable if it is easy to implement and does not violate known constraints. An idea is relevant if it satisfies pre-defined goals. An idea is specific if it has been worked out in detail.

Comparable sets of sub-dimensions have been suggested in the user innovation literature. Kristensson, Gustafsson and Archer (2004) compared the ideation performance of ordinary users, expert users and professionals. They used three quality attributes: originality (comparable to the novelty dimension suggested by Dean et al. (2006)), realisability (comparable to the feasibility dimension) and value (comparable to the relevance dimension). In a similar study, Magnusson (2009) compared the ideation performance of professionals, technically skilled users, ordinary users, consulting users and creativity-trained ordinary users. He used the quality attributes originality (comparable to novelty), producibility (comparable to feasibility) and user-value (comparable to relevance). Using the same attributes, Magnusson et al. (2014) compared technically skilled users with technically naïve users. Poetz and Schreier (2012) compared the ideas of users and professionals in terms of the attributes novelty, feasibility and customer benefit (comparable to value). Based on the four studies that have a product-user ideation focus, we chose novelty, feasibility and value as the quality attributes for our study.

Procedure

We established contact with Norwegian craft brewery *Nøgne Ø*. The brewery was founded in 2002 by two Norwegian home brewers and is nowadays part of Norwegian brewery group Hansa Borg Bryggerier. In 2015, *Nøgne Ø* produced 30 different styles of ales and exported to more than 40 markets. Two company professionals were recruited as expert raters. Expert 1 was 29 years old, female and had a business school background. Her responsibilities at *Nøgne Ø* were sales and logistics. At the time the study was conducted, she had been working for the brewery for 12 years. Expert 2 was 40 years old, male and had an engineering background. His responsibilities at *Nøgne Ø* were related to marketing and the web shop. At the time the study was conducted, he had been working for the brewery for 4.5 years.

The experts evaluated the 200 texts one-by-one and independently from each other. First, the experts were instructed to read the respective text carefully. Then, they were asked: *"Please evaluate if you think that the text contains one or more ideas"* and to respond on a binary *"yes"* versus *"no"* scale. If the expert responded *"yes"*, three rating scales were presented on which the expert was asked to evaluate the quality of the idea in terms of the three attributes novelty, feasibility and value. The scales were horizontally aligned ranging from very low (1) to very high (10). The instruction for the novelty attribute was: *"Please evaluate the novelty of the idea(s) in the text (by this we mean: to what degree does the idea suggest something new)"*. The instruction for the feasibility attribute was: "*Pleas evaluate the feasibility of the idea(s) in the text (by this we mean: to what degree is it possible to implement the idea)"*. The instruction for the value attribute was: *"Please evaluate the value of the idea(s) in the text (by this we mean: to what degree does the idea solve the underlying problem)"*.

Inter-rater reliability

To assess the inter-rater reliability of the idea/non-idea classification task, we calculated Cohen's kappa, normalised for differences between raters in their marginal distributions. The normalised version of kappa takes on values between 0 and 1 where a value of 0 stands for chance-level agreement and a value of 1 for the theoretical maximum of agreement, given the marginal distributions of the raters. Expert 1 identified 41 texts as containing ideas and 159 as not containing ideas. Expert 2 identified 87 texts as containing ideas and 113 as not containing ideas. They agreed on 35 texts as containing ideas and 107 as not containing ideas (See Table 2 for examples). These counts correspond to a normalised kappa of 0.74, suggesting that there was substantial agreement between the two experts as to whether a given text did or did not contain an idea (Cohen, 1960; Landis & Koch, 1977; von Eye & von Eye, 2008).

--- Table 2 ---

To assess the inter-rater reliability of the idea quality rating task, we calculated reliability measures based on generalisability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001). Only the 69 texts which the machine learning classifier had classified as an idea and which at least one of the brewery professionals had identified as an idea were included in the analysis. The design was a two-facet crossed design with tasks (the three quality attributes) and raters (the two brewery professionals) treated as fixed effects. The reliability (generalisability coefficient) of the averaged rating of a randomly picked idea text on the three attributes by the two raters was $E\rho^2 = 0.71$.

Results

Presence of ideas

Since our two company professionals had not perfectly agreed with each other on the presence or absence of ideas in the texts, we defined two validation criteria: a lenient criterion (Boolean OR: at least one professional had identified the respective text as containing an idea) and a strict criterion (Boolean AND: both professionals had identified the respective text as containing an idea).

Using the lenient criterion as a gold standard (where 47% of the 200 texts would be defined as true idea texts), the automatic idea detection system performed well. The classifier agreed with the company professionals in 77% of the cases as to whether a text did or did not contain an idea (accuracy). 75% of the texts which the classifier had identified as idea texts were also identified as idea texts by the company professionals (precision, also referred to as positive predictive value in the literature). The classifier correctly identified as idea texts 74% of the texts the professionals had identified as ideas (recall, also referred to as sensitivity or true positive rate in the literature). Since precision and recall always represent a trade-off, we also calculated their harmonic mean, the F_1 measure, as a compromise. Using the lenient criterion, it reached a very respectable value of $F_1 = 0.75$. For comparison see Christensen, Nørskov, et al. (2017) who obtained $F_1 = 0.54$, $F_1 = 0.55$ and $F_1 = 0.81$. Classification accuracy statistics are reported in Table 3.

Using the strict criterion as a gold standard (where only 18% of the 200 texts would be defined as containing ideas), the automatic idea classification system still agreed with the company professionals in 67% of the cases as to whether a text did or did not contain an idea (accuracy). Due to the much stricter criterion as to what defined an idea text, the precision of the classifier was lower: only 33% of the texts which the classifier had identified as idea texts were also identified as idea texts by the company professionals. For the same reason, recall was higher: the classifier correctly identified as idea texts 86% of the texts the professionals had identified as ideas. The F_1 measure, as a compromise between precision and recall, reached a value of 0.47.

Taken together, the criterion validity of the automatic idea detection system can be regarded as satisfactory as long as it is used for the screening of potential ideas. Deployed in a company as a tool for filtering out candidate ideas for product and process innovations, it may significantly reduce the time and effort that would otherwise have to be spent by company staff on manual screening and preliminary evaluation of a number of user contributions in potentially relevant online fora.

--- Table 3 ---

Quality of automatically detected ideas

Figure 2 shows the distribution of the quality ratings of the ideas (i.e., those texts that had been identified as ideas by the automatic idea detection system and which had been also been identified as ideas by at least one of the two company professionals). For texts which both company professionals had classified as an idea, the values on the novelty, feasibility and value attributes are the averaged ratings of both company professionals. For texts which only one of the company professionals had identified as an idea, the values are the ratings given by that professional. The overall quality values were calculated as unweighted averages of the ratings on the novelty, feasibility and value attributes.

--- Figure 2 ---

The distribution of the novelty ratings was concentrated in the lower range of the response scale (which had a minimum of 1 and a maximum of 10), the distribution of the feasibility ratings in the upper range of the response scale, and the distributions of the value ratings and overall quality in the middle of the response scale. The results suggest that, on average, the ideas which the automatic idea detection system extracted from the *alt.beer.home-brewing* community appeared rather feasible to brewery professionals, were not particularly novel, but had medium value and medium overall idea quality. Although the results are generally consistent with the findings of Kristensson et al. (2004), Magnusson (2009), Poetz and Schreier (2012) and Magnusson et al. (2014), the algorithmically identified ideas in the present study were on average slightly less novel but also slightly more feasible than the ideas generated by the human users who directly participated in the above-cited studies.

Discussion and conclusion

Implications for researchers and practitioners

The age of "big data" has generated opportunities and challenges for companies. In the present study, we focused on the case of data generated by users of social media and online communities, which can pose a challenge due the semi- or unstructured nature of such data (Olsen & Christensen, 2015). If people's thoughts and ideas expressed on social media can be captured in a systematic manner, inbound open innovation processes can be accelerated (McAfee & Brynjolfsson, 2012; Jin, Wah, Cheng, & Wang, 2015) and thereby make companies more receptive (Ooms et al., 2015). The first aim of the present study was to investigate if ideas for product- and process innovations detected by an artificial intelligence system (in this case, the one developed by Christensen, Nørskov, et al. (2017) would also be regarded as ideas by company-internal staff who would be responsible for taking the ideas further in the innovation process.

Our results suggest that this is to a considerable extent the case: the performance of the system can be regarded as sufficient for an initial screening of potential ideas. Deployed in a company as a tool for selecting candidate ideas for product and process innovations, it can significantly reduce the time and effort that would otherwise have to be spent by company staff on wading through a large number of user contributions in potentially relevant online communities. The exact level of criterion-related validity that our system could achieve depended on several factors. The most important of these are (a) the definition of the "gold standard" against which the predictions are validated and (b) the cut-off used for transforming the continuous posterior probability score generated by the system into a binary prediction. In our analysis, we used two of the possible gold standards: a lenient criterion (at least one of the company professionals had rated the respective text as containing an idea) and a strict criterion (both company professionals had rated the text as containing an idea). The lenient criterion led to an implied base rate of 47% for the target event (i.e., the probability that a randomly chosen text from among the 200 used in the present study would contain an idea), whereas the strict criterion reduced the implied base rate to 18%. It is not possible to define on purely statistical grounds what the right base rate should be. This is complicated by the fact that the two company professionals who served as experts in our study did not have the same base rates in their individual classifications: Expert 1 appeared to use a more conservative standard of judgment, rating 21% of the 200 texts as containing ideas, whilst Expert 2 appeared to use a more liberal standard, rating 44% of the texts as containing ideas.

Whether it makes more sense for a given company to use a stricter or more lenient criterion for further filtering of the automatically identified ideas may depend more on strategy and available resources: a lenient criterion may be more appropriate if a company wants to cast its net wide and thereby reduce the risk of missing certain ideas which might not yet be able to achieve full cross-functional consensus. However, the company would also have to be prepared to assign the necessary resources for dealing with the larger number of ideas that would enter the innovation funnel. If, on the other hand, a company wants to limit its resource expenditure and focus on ideas that can already in the early phases achieve cross-functional consensus, a stricter criterion would be appropriate.

A similar objective can be achieved by tuning the cut-off value of the SVM classifier underlying the Christensen et al. (2017) system. The algorithm yields a posterior probability score that is continuous on the (0,1) interval. A traditional way of transforming the posterior probability score into a binary classification rule is to use the value 0.50 as a cut-off such that a text is classified as an idea text if the probability that the text contains an idea, given the support vectors, is larger than 0.50, and classified as a non-idea text otherwise. However, the traditional way of setting the cut-off may not always be the most useful way. Another heuristic that is typically more useful is to set the cut-off equal to one minus the base rate of the target even, either on the posterior probability scale or on the empirical percentile scale. This heuristic would match the prior probability of classifying a text as an idea to the base rate of the event. A third way of setting the cut-off is to estimate how many additional ideas a company would be able to absorb into its innovation funnel and to use an appropriate absolute cut-off, selecting the right number of ideas from the top of the posterior probability ranking.

The second aim of the present study was to investigate if the automatic idea detection system developed by Christensen, Nørskov et al. (2017) would extract *good* ideas from the online community that served as an example here. For the online community under investigation, our answer is a qualified yes: the distribution of the overall idea quality score, calculated as the average rating of each idea on the three quality attributes (novelty, feasibility, value) by the two company professionals, was concentrated in the middle of the response scale (mean = 4.8, 25^{th} percentile = 3.8, 50^{th} percentile = 5, 75^{th} percentile = 5.7) and ranged from a minimum of 1 (the lower end of the response scale) to a maximum of 8 (two points below the maximum of the response scale). Overall, the ideas extracted by the automatic detection system appear to have made a reasonable impression on the company professionals.

Another important aspect for the evaluation of innovative ideas is their *timing*. Although we did not explicitly focus on this aspect in the present study, many of the ideas identified by our system could serve as good cases here. Take the idea about gluten-free beer in Table 2 as an example. It was identified as an idea by the automatic idea detection system and by both company professionals. Notably, the idea was posted to the community in 2005, one year before the world's first gluten-free beer was launched ("New Grist" by *Lakefront Brewery Inc.*, Milwaukee, WI, launched in 2006). An interesting avenue for future research would be to follow up more systematically on how often, and with which lead time, user-contributed ideas in online communities precede the development and market launch of commercial products.

An interesting detail related to the quality evaluations is that the identified ideas tended to be regarded as more feasible and valuable by our company professionals than they were regarded as novel. This finding reflects results obtained by Kristensson et al. (2004). However, as already observed, agreement between our experts was not perfect here either. As an example, consider the text shown in Table 4: a community member suggests a new mead recipe. Overall, the idea was rated as one of the best by the two company professionals. Expert 1 assigned a rating of 2 on the novelty attribute, 7 on feasibility and 4 on value. Expert 2 rated it 9 on novelty, 9 on feasibility and 9 on value. In the additional, qualitative responses we obtained from the two professionals, it became clear that Expert 1 evaluated the idea in terms of its quality as an idea for process innovation whereas Expert 2 evaluated it in terms of its quality as idea for product innovation. Different perspectives, either due to the functional specialisation of our company professionals or due to their different levels of experience with the product category, seem to have led to different standards of judgment.

--- Table 4 ---

Limitations and suggestions for future research

The results presented here are an evaluation of a particular automatic idea detection system (the one developed by Christensen, Nørskov et al., 2017) to a particular case (the craft brewing community *alt.beer.home-brewing*), evaluated from the point of view of two brewing professionals connected to a particular craft brewing company (*Nøgne Ø*). Naturally, this poses limits to the generalisability of our findings. The ideas detected by an automated system can only be as good as the ideas voiced by the users in the online community under investigation. Furthermore, the 200 texts we selected for evaluation were only a sample and therefore unlikely to reflect the whole range of ideas discussed in the community. It is an open question whether similar results will be achieved when automatic idea detection systems are applied to other technology domains or product categories.

This question can only be answered by follow-up research. However, we do believe that we have demonstrated the potential of automatic idea identification systems: they can be a powerful technique for the harvesting and initial screening of user ideas from online fora that do not conform, and are not limited to, the highly restrictive architecture and user basis of dedicated crowdsourcing systems. We hope that studies such as ours can also also make a contribution to a wider discussion: which business tasks of a more complex nature can credibly be solved by artificial intelligence-based systems? We are convinced that the answer does not only lie in what is technically possible but also in what is acceptable to the prospective users of the information generated by such systems. More user evaluations of the performance of artificial intelligence-based systems are needed.

The presented method adds a new channel for feeding ideas into the innovation processes of firms, complementing company-hosted idea crowdsourcing communities such as those studied by di Gangi et al. (2010) and Poetz and Schreier (2012). But can the method completely substitute company-hosted idea crowdsourcing communities? In our opinion, the answer to this question is a qualified “no”. A company-hosted crowdsourcing community generates more than just ideas for the company. It also serves as an arena for cultivating customer relations, enabling the company to interact directly with its most dedicated customers. The method evaluated in this paper is strictly “one-way” and does not offer such opportunities. The method can, however, dramatically reduce the costs of crowdsourcing new ideas, which is particularly relevant for companies who do not enjoy such large and loyal customer bases as Dell (di Gangi et al., 2010) or Lego (Antorini et al., 2012). This we see as a key advantage of the method, and it may even level the competitive playing field between large companies and SMEs.

References

Antons, D., Kleer, R., & Salge, T. O. (2016). Mapping the Topic Landscape of *JPIM* , 1984-2013: In Search of Hidden Structures and Development Trajectories: Mapping the Topic Landscape of *JPIM* , 1984-2013. *Journal of Product Innovation Management*, *33*(6), 726–749.

Antorini, Y. M. (2007). *Brand Community Innovation: An Intrinsic Case Study of the Adult Fans of LEGO Community*. Copenhagen Business School, Frederiksberg: Center for Europaforskning,.

Antorini, Y. M., Muñiz, J., Albert M., & Askildsen, T. (2012). Collaborating With Customer Communities: Lessons from the Lego Group. *MIT Sloan Management Review*, *53*(3), 73–95.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Christensen, K., Liland, K. H., Kvall, K., Risvik, E., Biancolillo, A., Scholderer, J., … Næs, T. (2017). Mining online community data: The nature of ideas in online communities. *Food Quality and Preference*, *62*, 246–256.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine and text mining. *Creativity and Innovation Management*, *26*(1), 17–30.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Dean, D. L., Hender, J. M., Rodgers, T. L., & Santanen, E. L. (2006). Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, *7*(1), 646–698.

di Gangi, P. M., Wasko, M. M., & Hooker, R. E. (2010). Getting Customers' Ideas To Work For You: Learning from Dell How To Succeed With Online User Innovation Communities. *MIS Quarterly Executive*, *9*(4), 213–228.

Ekvall, G. (1997). Organizational conditions and levels of creativity. *Creativity and Innovation Management*, *6*(4), 11.

Frederiksen, M. H., & Knudsen, M. P. (2017). From Creative Ideas to Innovation Performance: The Role of Assessment Criteria. *Creativity and Innovation Management*, *26*(1), 60–74.

Füller, J., Bartl, M., Ernst, H., & Mühlbacher, H. (2006). Community based innovation: How to integrate members of virtual communities into new product development. *Electronic Commerce Research*, *6*(1), 57–73.

Füller, J., Jawecki, G., & Mühlbacher, H. (2007). Innovation creation by online basketball communities. *Journal of Business Research*, *60*(1), 60–71.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Hsinchun Chen, Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188.

Jeppesen, L. B., & Frederiksen, L. (2006). Why Do Users Contribute to Firm-Hosted User Communities? The Case of Computer-Controlled Music Instruments. *Organization Science*, *17*(1), 45–63.

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. *Big Data Research*, *2*(2), 59–64.

Kristensson, P., Gustafsson, A., & Archer, T. (2004). Harnessing the Creative Potential among Users*. *Journal of Product Innovation Management*, *21*(1), 4–14.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, *52*(2), 481–495.

Magnusson, P. R. (2009). Exploring the Contributions of Involving Ordinary Users in Ideation of Technology-Based Services*. *Journal of Product Innovation Management*, *26*(5), 578–593.

Magnusson, P. R., Wästlund, E., & Netz, J. (2014). Exploring Users' Appropriateness as a Proxy for Experts When Screening New Product/Service Ideas*. *Journal of Product Innovation Management*, *33*(1), 4–18.

Mahr, D., & Lievens, A. (2012). Virtual lead user communities: Drivers of knowledge creation for innovation. *Research Policy*, *41*(1), 167–177.

Majchrzak, A., & Malhotra, A. (2013). Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems*, *22*(4), 257–268.

McAfee, A., & Brynjolfsson, E. (2012). Big data. The management revolution. *Harvard Business Review*, (October), 3–9.

Nørskov, S., Antorini, Y. M., & Jensen, M. B. (2015). Innovative brand community members and their willingness to share ideas with companies. *International Journal of Innovation Management*.

Nunan, D., & Di Domenico, M. (2013). Market research and the ethics of big data. *International Journal of Market Research*, *55*(4), 2–13.

Olsen, N. V., & Christensen, K. (2015). Social media, new digital technologies and their potential application in sensory and consumer research. *Current Opinion in Food Science*, *3*, 23–26.

Ooms, W., Bell, J., & Kok, R. A. W. (2015). Use of Social Media in Inbound Open Innovation: Building Capabilities for Absorptive Capacity: Use of Social Media in Inbound Open Innovation. *Creativity and Innovation Management*, *24*(1), 136–150.

Poetz, M. K., & Schreier, M. (2012). The value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas?*. *Journal of Product Innovation Management*, *29*(2), 245–256.

Thorleuchter, D., & Van den Poel, D. (2013). Web mining based extraction of problem solution ideas. *Expert Systems with Applications*, *40*(10), 3961–3969.

Van de Ven, A. (1986). Central Problems in the Management of Innovation. *Management Science*, *32*(5), 590–607.

van den Ende, J., Frederiksen, L., & Prencipe, A. (2015). The Front End of Innovation: Organizing Search for Ideas. *Journal of Product Innovation Management*, *32*(4), 482–487.

Vandenbosch, B., Saatcioglu, A., & Fay, S. (2006). Idea management: a systemic view. *Journal of Management Studies*, *43*(2), 259–288.

von Eye, A., & von Eye, M. (2008). On the marginal dependency of Cohen's κ. *European Psychologist*, *13*(4), 305–315.

von Hippel, E., Ogawa, S., & PJ de Jong, J. (2011). The age of the consumer-innovator.

von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, *32*(7), 1217–1241.

von Krogh, G., & von Hippel, E. (2006). The Promise of Research on Open Source Software. *Management Science*, *52*(7), 975–983.

Zwitter, A. (2014). Big Data ethics. *Big Data & Society*, *1*(2).

Table 1 - Performance of the automatic idea detection system

Partition	True positives (TP)	True negatives (TN)	False positives (FP)	False negatives (FN)	Classification accuracy	Precision	Recall	F_1
Validation set	27%	70%	1%	2%	0.97	0.97	0.92	0.94
Hold-out set	25%	70%	1%	3%	0.96	0.96	0.88	0.92

Table 2 - Example of an idea text and a non-idea text on which both raters agreed

Idea text	Non-idea text
´Buckwheat has been used as an adjunct for a long time in a few beers. It also is used to make gluten free beers. It has a high gelatinisation temp so need to be boiled first. Extract potential is about 1.032. Can be used lightly roasted to add colour to gluten free beers, or use Kasha (a roasted buchwheat). I think Rogues make a buckwheat ale´	´Thanks for the help. My internet is screwy or I would have replied sooner. I re- pitched and it is going crazy. A load off my mind! now i can concentrate on getting another cider and a wit going. Anyone have any suggestions for a good belgian style ale like duvel? I am an extract with specialty grains level brewer, so whole grain is out for now. Thanks again for all the help!´

Table 3 - Presence of ideas: classification accuracy of the automatic idea detection system, validated against the judgments of two company professionals

Validation criterion	True positives (TP)	True negatives (TN)	False positives (FP)	False negatives (FN)	Classification accuracy	Precision	Recall	F_1
Lenient criterion:								
Classified as idea by Expert 1 OR Expert 2	35%	42%	12%	12%	0.77	0.75	0.74	0.75
Strict criterion:								
Classified as idea by Expert 1 AND Expert 2	15%	52%	31%	3%	0.67	0.33	0.86	0.47

Table 4 - Idea text identified by classifier, Expert 1 and Expert 2

´I've made several batches. Below is my recipe The love of my life I love Mead as you can probably tell. Please note, this is Mead but I do not use any water. I use apple juice as the base. You can use water but I find the apple juice makes it a bit nicer for those of you who love apples and like a high alcohol content. No citric acid needed. This is called Apple Honey Melonomel Meade You will need... 1 Package Red Star wine yeast 4 Gallons apple juice from concentrate 2-5 pounds of pure honey, the more the better. This shit is expensive though. 1 cup table sugar 5 Fuji apples Siphen hose, any small tube will work. A 5 gallon carboy or tub 1 balloon Step one, crush your apples or use a blender. Step two, boil apples in large pot with apple juice. Step three, set aside to cool Step Four, boil honey in large pot of apple juice Step five, set aside to cool. Step six, dump mixture into large 5 gallon carboy and add activated yeast. Step six, allow the mead to ferment for 3-4 weeks, once fermentation begins to slow prime with table sugar by dilluting the 1 cup of table sugar in 1/2 gallon of apple juice then pour this directly into the carboy. A balloon can be placed over the mouth of the carboy to monitor the fermentation. Simply peirce a small hole in the baloon to allow CO2 to escape. Once the Meade has cleared (meaning you can read a newspaper through it) transfer it into a secondary (Save the sediment for use as the Yeast in your next batch of Meade) and let it clarify for 2-3 weeks. After this bottle the meade and let fermintation finish off. Total process about 70 days and its ready to drink. This will burn going down but is smooth as a whistle. Enjoy....´

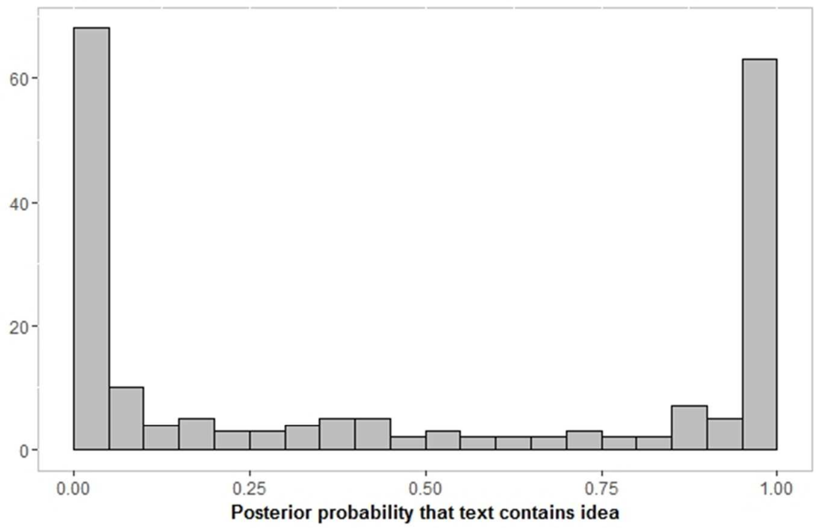

Figure 1 - Histogram of the posterior probability scores generated by the SVM-based automatic idea detection system for the 200 texts used in the present study

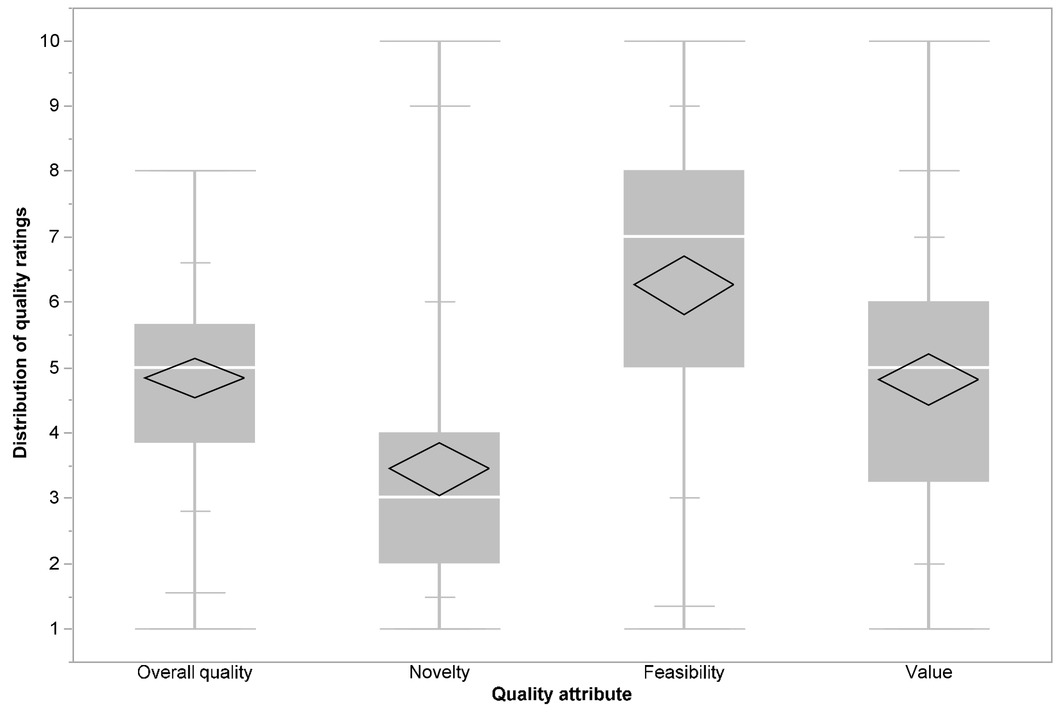

Figure 2 - Box plots of the distribution of quality ratings (overall quality = unweighted average of novelty, feasibility and value; diamonds represent 95% confidence intervals around distribution means)