Detecting ideas in online communities

Utilizing machine learning and text mining for finding ideas in online communities

Identifisering av ideer i nettsamfunn

Utnyttelse av maskinlæring og tekstmining til å finne ideer i nettsamfunn

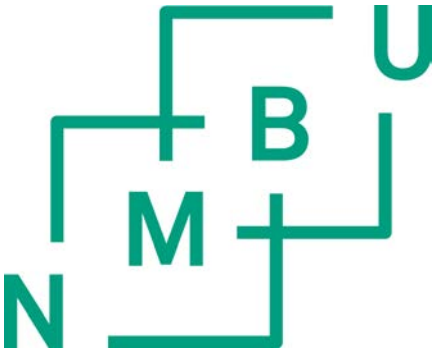
Philosophiae Doctor (PhD) Thesis

Kasper Knoblauch Christensen

Faculty of Science and Technology

Norwegian University of Life Sciences

Ås 2017



Thesis number 2017:18
ISSN 1894-6402
ISBN 978-82-575-1424-2

Supervisors:

Professor **Knut Kvaal**, Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Dr. Argic **Einar Risvik**, Senior Research Scientist, Department of Sensory, Consumer and Innovation, Nofima, Ås, Norway

Professor **Tormod Næs**, Senior Research Scientist, Nofima, Ås, Norway *and* Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Copenhagen, Denmark

Dr. **Torulf Mollestad**, Principal consultant, Altran, Norway, Oslo

Evaluation committee:

Professor **Per B. Brockhoff**, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

Dr. **Hal Macfie**, Visiting Professor, Universities of Reading, Nottingham, United Kingdom

Associate Professor, **Jorge M. Marchetti**, Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Detecting ideas in online communities: Utilizing machine learning and text mining for finding ideas in online communities

PhD Thesis · 2017 © Kasper Knoblauch Christensen
ISBN: 978-82-575-1424-2

Conor O'Brien "Running out of ideas" retrieved from www.conorobrienart.com

Acknowledgements

I would like to thank the Norwegian University of life sciences (NMBU) for giving me the opportunity to study my doctoral degree at their university. It has given me the opportunity to work with many great scientists and kind people. Knut Kvaal, Einar Risvik, Tormod Næs and Torulf Mollestad have provided me with the best possible supervision I could have hoped for. Joachim Scholderer, Lars Frederiksen and Nina Veflen also deserve to be mentioned in this regard. I owe all of them much recognition for having helped me in the process from being a master student to becoming a PhD student.

The four scientific papers and manuscripts included in this work, gave me the opportunity to work with other people than my team of supervisors. Sladjana Nørskov has been a great help and her knowledge about online communities has been extremely useful. Alessandra Biancolillo helped me understand the technicalities of Partial Least Squares and Kristian Hovde Liland has been a fantastic help because of his strong programming skills and knowledge about informatics and statistics. I am also grateful for having had extremely good colleagues at Nofima. Even though many of these colleagues have not had a direct role to play in my project, they were a central part of the social life at Nofima and at Ås. Margrethe, Øydis, Paula, Karen, Marta, Jens Erik, Stine, Antje, Ida and Sveinung are all persons who I owe thanks you for their kindness and openness.

Last but not least, I would like to thank the "*The foundation for research and levy on agricultural products*" in Norway as well as Nofima for funding my project. I have learned a lot from the PhD process and I hope that society and science will benefit from the results.

Kasper K. Christensen,

Ås, January 2017

Abstract

Online communities serve as a gathering point for dedicated product users and consumers who discuss all imaginable topics. Scholars have argued that this discussion can lead to new ideas useful for firms. That is, if the ideas can be detected amongst the vast amount of information contained in online communities. The nature of online community data makes idea detection labour intensive and a systematic way of dealing with the data is needed if firms are to fully exploit online community ideas for innovation. This is the starting point for the research carried out in this doctoral project.

The present doctoral thesis introduces an automatic method for idea detection aimed at screening large amounts of online community texts. The method is based on machine learning and text mining techniques and it is developed on two product cases related to brewing and Lego. The method relies on a large set of pre detected idea texts and non-idea texts for learning the lexical pattern embedded in idea texts. It is described how to pre-process the text data and how to adjust the machine learning techniques for optimal idea detection performance. Support Vector Machines and Partial Least Squares are used as machine learning techniques.

The presented results show that when the method is trained for Lego idea detection and tested on an independent Lego hold-out set, the method obtains moderate to substantial agreement with human idea raters. When the method is trained for beer brewing idea detection on an independent hold-out set, the method obtains fair to substantial agreement with two brewing experts. Moreover the results indicate that people use specific idea words and expressions when they talk about ideas. This is why automatic idea detection is possible.

Sammendrag

Nettsamfunn er et samlingspunkt for dedikerte produktbrukere og forbrukere som diskuterer ethvert tenkelig emne. Forskere har hevdet at denne diskusjonen kan føre til nye ideer som er nyttige for bedrifter. Det er hvis, ideer kan bli identifisert blant de store mengder data nettsamfunn genererer. Naturen av data i nettsamfunn gjør ideidentifikasjon arbeidskrevende og dette må håndteres systematisk dersom bedrifter ønsker å dra nytte av ideer fra nettsamfunn for innovasjon. Dette er utgangspunktet for forskningen som er gjennomført i dette doktorgradsprosjektet.

Denne avhandlingen presenterer en automatisert metode for idéidentifikation for å muliggjøre en filtrering av tusenvis av tekster fra nettsamfunn. Metoden er basert på maskinlæring og tekstmining. Metoden er utviklet på to produktgrupper knyttet til brygging av øl og Lego. Den er basert på trening av algoritmer med hensyn til et stort sett med pre-identifiserte idétekster og ikke-ideetekster. Algoritmene brukes for å lære en teknikk å identifiserer de syntaktiske mønstre i ideetekster. Det beskrives hvordan teksten preprosesseres, og hvordan en justerer maskinlærings teknikker for optimal idé identifikasjon og ytelse. Support Vector Machines og Partial Least Squares ble brukt som maskinlærings teknikker.

De presenterte resultatene viser at når algoritmene er trent til Lego ideidentifikasjon og testet på et uavhengig Lego idesæt, oppnås en moderat til betydelig overensstemmelse med dommerne brukt i studiet. Når metoden er trent til øl-ideidentifikasjon på et uavhengig øl-datasett oppnås er det betydelig samsvar med to bryggeri-eksperters bedømmelse. Videre viser resultatene at folk bruker bestemte idéord og uttrykk når de snakker ideer. Dette er årsaken til automatisk ideidentifisering er mulig.

Table of contents

List of papers

1. Kasper Christensen, Sladjana Nørskov, Lars Frederiksen Joachim Scholderer. *In search of new product ideas: Identifying ideas in online communities by machine learning and text mining.* Creativity and Innovation Management Journal (In press). 2016

2. Nina Veflen Olsen, Kasper Christensen. *Social media, new digital technologies and their potential application in sensory and consumer research.* Current Opinion in Food Science, 3, 23-26. 2015

3. Kasper Christensen, Knut Kvaal, Einar Risvik, Alessandra Biancolillo, Kristian Hovde Liland, Joachim Scholderer, Sladjana Nørskov, Tormod Næs. *Mining online community data: The nature of ideas in online communities.* (Submitted for publication in Food Quality and Preference)

4. Kasper Christensen, Joachim Scholderer, Stine Alm Hersleth, Tormod Næs, Knut Kvaal, Torulf Mollestad, Nina Veflen, Einar Risvik. *How good are ideas identified by an automatic idea detection system?* (Submitted for publication in Creativity and Innovation Management Journal)

Introduction

"All innovations originate from ideas" (Boeddrich, 2004, p. 274,). Ideas represent the sparks that ignites the innovation engine that drives the development of firms and society. Probably this is why *"The front end of innovation: Organizing search for ideas"* was the theme for a special issue in "The Journal of Product Innovation Management" in 2014 (van den Ende, Frederiksen, & Prencipe, 2015). The present doctoral thesis addresses how firms can *search* for ideas amongst the vast amount of data contained in online communities so that innovation can be accelerated.

There is already great deal of research on the process of *generating* ideas and *evaluating* quality of ideas for new products. See e.g. Dean, Hender, Rodgers, & Santanen (2006). In many studies the sources to ideas, are often the employees of the firm or the users of the firms products (Kristensson, Gustafsson, & Archer, 2004; Magnusson, 2009; di Gangi, Wasko, & Hooker, 2010; Soukhoroukova, Spann, & Skiera, 2012; Poetz & Schreier, 2012; Bayus, 2013). As information technology has developed, and society has become more digitized, new places where ideas emerge have arisen. Now ideas also emerge outside the boundaries of the firm in virtual places on the internet referred to as *online communities.* As a consequence of this, online communities have caught the interest of scholars as well as practitioners. The central driver for this increased attention is that online communities allow geographically dispersed people to interact and develop and share knowledge. The outcome of this interaction is *new* knowledge and *new* ideas that may be useful for innovation (Lee & Cole, 2003; Jeppesen & Frederiksen, 2006, Füller, Bartl, Ernst, & Mühlbacher, 2006; Füller, Jawecki, & Mühlbacher, 2007; Dahlander, Frederiksen, & Rullani, 2008; Antorini, Muñiz, & Askildsen, 2012; Nørskov, Antorini, & Jensen, 2015). Online communities constitute a new type of arena for knowledge generation and ideation, and they demand new methods that

enable researchers as well as practitioners to analyze the information they contain. This challenge must be overcome if firms and society are to utilize the potential online communities possess for innovation.

The overall aim of this doctoral thesis is to investigate if it is possible to automatically detect ideas written in online communities via a type of artificial intelligence system based on machine learning and text mining. *Paper 1* provides a proof of concept. It aims to investigate if it is possible to use machine learning and text mining to automatically detect ideas. A Lego community is used as case. *Paper 2* is an outlook paper. It discusses and highlights the potential of new digital technologies for food science. *Paper 3* investigates the textual pattern that makes automatic idea detection possible. As a central part the paper addresses and analyses the words and expressions online community members use when they express ideas. A community related to Lego is used as case and a community related to brewing is used a case. In *Paper 4* it is tested if firm employees consider the ideas detected by the automatic system, *good* ideas that can potentially become innovations. Figure 1 displays the sequential order of the papers included in this work.

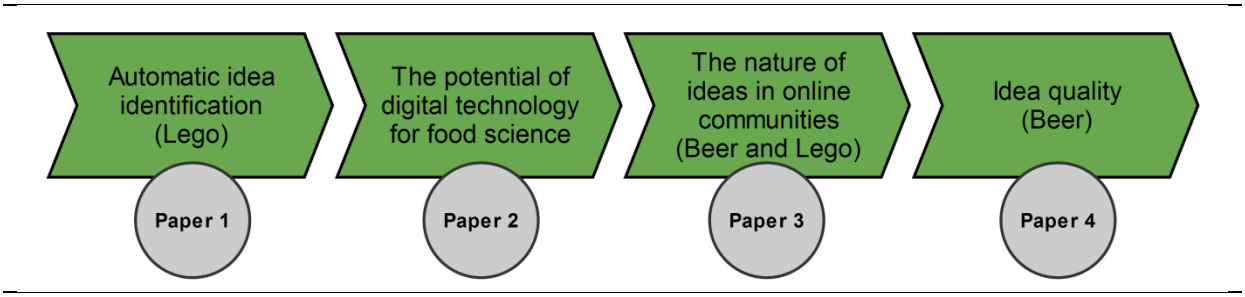


Figure 1 – Sequential order of papers included in thesis

Theoretical background and Aims

Innovation: From an idea to an innovation

The concept of innovation is confuse, and to understand why ideas are relevant, it is central to understand what innovation *is* and how innovation is related to the development of firms and society. In an attempt to reach clarity about the concept of innovation, Baregheh, Rowley, & Sambrook (2009) identified 60 innovation definitions already published in the literature. They conclude that: *"Innovation is the multi-stage process whereby organizations transform ideas into new/improved products, services and processes, in order to advance, compete and differentiate themselves successfully in the marketplace."* (Baregheh et al., 2009, p. 1334, line 1-3). In a similar manner the *process* of innovation can be defined as: *"the development and implementation of new ideas by people who over time engage in transactions with others within an institutional context"* (Van de Ven, 1986; p. 3, line 12-14; Björk & Magnusson, 2009). This is in line with Schumpeter (1943) who suggest that innovation can be seen as a continuous process.

Both definitions stated above are relevant because they suggest that innovation does *not* call for a specific type of outcome within a specific product category. This means that the innovation process does not need to results in a tablet, an electric car, a cure for cancer, a healthy burger or bigger-, better-, tastier- and fresher salmon. Neither do the definitions state that innovation process calls for any monetary output. The definitions state that the innovation process is the development of new ideas that can be implemented and it fits to *all* types of product categories. Depending on the product category, the innovation process may then transform into the development of for example a new food product that should be *implemented*. Here implementation means that not only should the new product be *developed*. The new product should also be accepted by the *end-user* (the end-user is often, but not

necessarily, called the consumer and the two concepts can be considered synonyms for the remaining part of this thesis). The concept of implementation is central argument because a new product is not *accepted* by the end-user just because it is *developed*. It is only in the exact moment where the new product has been developed *and* accepted by the end-user, that the innovation process has resulted in *an innovation*. Therefore, the term *innovation* refers to the innovation *process* and the term "*an innovation*", refers to a tangible- or intangible product that has been *accepted* by the end-user. That is the difference between *innovation* and *an innovation*.

The innovation process

Scholars have argued that the innovation process consists of two phases. A planning phase and a development phase (Moenaert, De Meyer, Souder, & Deschoolmeester, 1995). The development phase can be divided into several steps or stages where the aim is to develop a new product that can be introduced to the market (i.e. the end-user). According to R. G. Cooper & Kleinschmidt (1986) the development phase in its most extensive form can be divided into 13 steps. These are: (1) Initial idea screening, (2) Preliminary market assessment, (3) Preliminary technical assessment, (4) Detailed marketing research, (5) Business analysis, (6) Product development, (7) In-house product testing, (8) Customer tests of product, (9) Trial sell, (10) Trial production, (11) Pre-commercialization business analysis, (12) Production start-up and (13) Market launch (See Figure 2 for illustration). In each of these steps, employees and managers operating within the firm, co-operate so that a given idea can be passed on to the next step before a given deadline. *In between* each step a go/kill decision is made by proper decision makers, for example a manager, and as the idea moves through the different steps, the idea matures and transforms into a product (Robert G. Cooper, 2008).

The process described above starts with the screening of ideas, so that the initial *best* idea can be selected and passed on to the preceding step. This suggests that without ideas

there would be nothing to use as start input for the innovation process. Therefore the next natural question is to ask: Where- and how can ideas be obtained so that the innovation process is constantly fed with new ideas? R. G. Cooper & Kleinschmidt (1986) writes that the ideas used in the screening phase, are typically generated based on information from sources from the market. For example market generated ideas can stem from sources such as salespersons, competitors and/or customers and consumers. This is unlike ideas that have been generated based on new opportunities that have risen because of technological advance. Ideas of this character are typically generated by in-house engineers who have learned about a new technology, but have little feeling with market demands. This may result in new products that have no market justification (See e.g. von Hippel (1986) and von Hippel & Foster (1988) for more on this topic).

It is not central whether ideas stem from the market or employees. What *is* central in relation to this thesis is that the planning phase comes *before* the development phase. In the planning phase ideas are generated and this phase is also known as the *Fuzzy Front-End* of innovation. The fuzzy front-end of innovation, refers to all the activities that comes *before* the first idea screening step (See Figure 2) (Smith & Reinertsen, 1991; Reid & De Brentani, 2004). It is where information is gathered and processed, with the aim of generating and/or developing an idea that is sufficiently good that it can be taken into consideration for further development. It is a phase that has high impact on the likelihood of success for the new product that is eventually developed (Kim & Wilemon, 2002). The more ideas that can be generated at the fuzzy front-end, and the better these ideas are, the higher is the likelihood the firm will succeed and useful product will be taken to market.

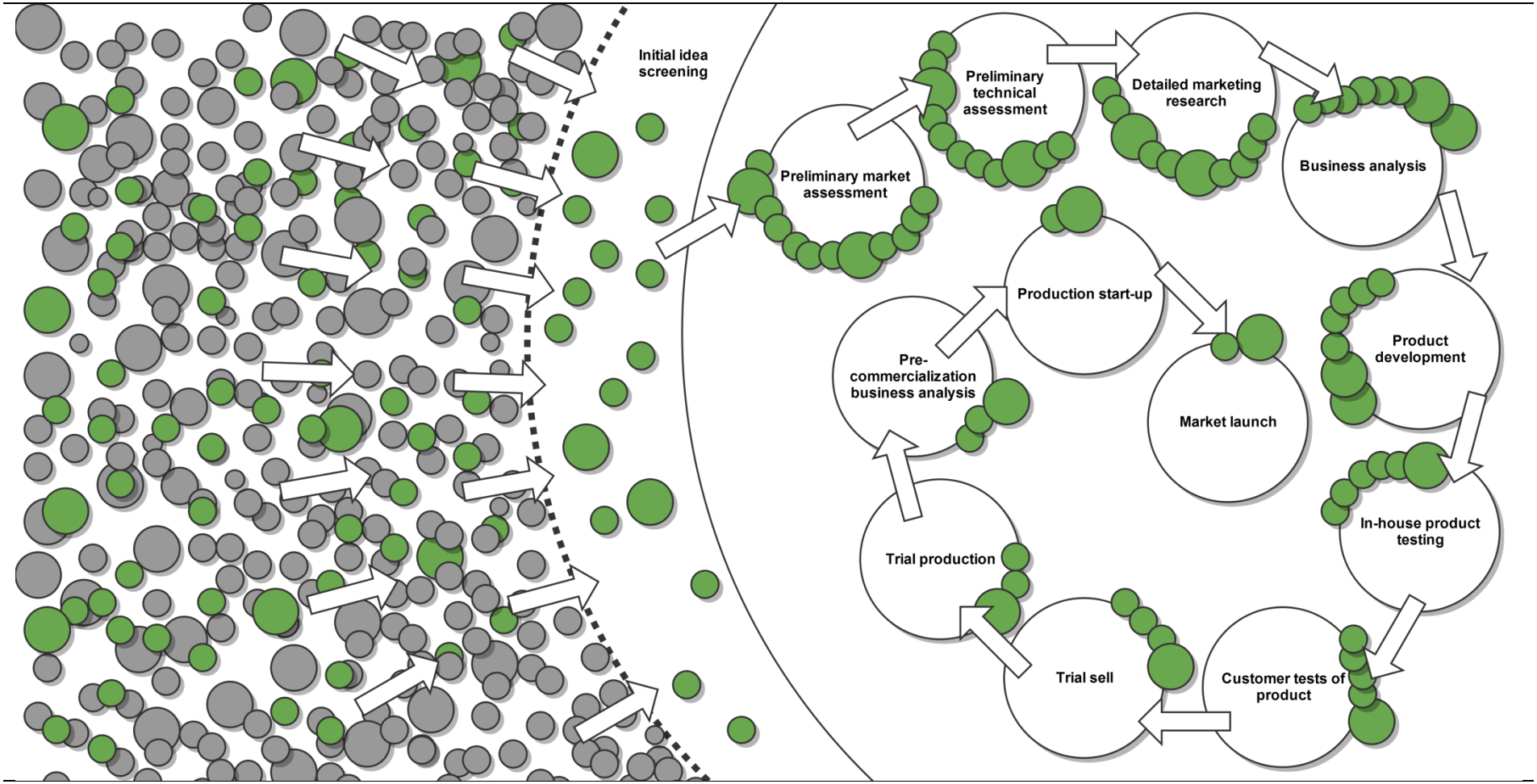


Figure 2 – Illustration of the innovation process from planning phase (Fuzzy Front-End) to development phase. The planning phase is to the left of the stippled line. The development phase is to the right of the stippled line. The grey bubbles illustrate information irrelevant for innovation. Green bubbles illustrate ideas that have survived the preceding step in the process of reaching market launch.

Online communities as idea reservoirs

Before the emergence of the internet, organizations like firms and universities were the primary drivers of knowledge generation and innovation for society. Now however, online communities can be seen as organizations equally important for knowledge generation. Online communities are a new potential source to ideas and knowledge (Dahlander et al., 2008; Dahlander & Magnusson, 2008). Firms, universities *and* online communities can all be seen as special types of organizations or places that allow people to collaborate and drive innovation forward (Lee & Cole, 2003). Online communities can be defined as: *"Groups of people with common interests and practices that communicate regularly and for some duration in an organized way over the internet through a common location or mechanism"* (Ridings, Gefen, & Arinze, 2002, p. 273, line 15-17). Many online communities are freely available to the public and the conversation is free for everyone to read and participate in. Facebook Groups and Google Groups are examples of digital places where people can gather and together constitute online communities.

One of the most prominent examples of how powerful online community innovation can be, is the case of *Open Source Software*. The concept of *software* is related to the development of computer code. Computer code can be seen as a large series of commands and instructions that controls our computers and makes them do what they do. Microsoft Windows and Office is probably the most well-known examples of firm-developed software, but also firms like for example the SAS institute and Apple develop software (von Hippel & von Krogh, Georg, 2003; Albors, Ramos, & Hervas, 2008). Open source software is an interesting case because the code is developed by programmers from all over the world who collaborate via online forums and/or mailing lists. They constitute their own online communities dedicated to develop useful computer code. The development typically happens for free because most of the programmers, who participate, spend their own time and effort

developing the code and software of pure intrinsic motivation. Online communities have provided the communication infrastructure and the meeting place for the development of this freely available software. Linux, R, Python and JavaScript are just a few examples of types of open source code. In this occasion it is appropriate to mention that all calculations and computations related to this thesis, was carried out via the programming language R in R-studio (R Core Team, 2017).

That online communities' can serve as a gathering point for dedicated computer programmers who freely contribute to innovation in software development, is just one example of online community based innovation. Many other online communities exist where people collaborate for idea generation and innovation within a given topic. The toy brick company Lego is one example of a firm who benefits from new product ideas stemming from online communities. In the case of Lego, thousands of Adult Fans of Lego, also called AFOL's, have gathered in online communities where they discuss Lego related topics. This discussion has led to, not only new product ideas, but also new business opportunities and exposure to new technologies (Antorini et al., 2012). Basketball shoes (Füller, Jawecki, & Mühlbacher, 2007), cars (Füller et al., 2006) and computers (di Gangi et al., 2010; Poetz & Schreier, 2012) are just a small amount of the product categories where online communities have proven useful for innovation.

Eric von Hippels *Lead-User* concept provides a central argument for why ideas stemming from online communities can be drivers for innovation (von Hippel & Foster, 1988). Von Hippel (1986) writes that a lead-user is a person who faces the same needs and demands as does the market. Here *the market* is represented by *ordinary* users who are typically *passive* consumers of a given product. For example Lego lead-users are people who enjoy building with Lego so much that that they spend time and effort thinking about what is the next Lego product they would want to acquire. The same logic applies for basketball

shoes, cars, computers, beers and many other product categories. Therefore, ideas stemming from online communities might not only be used to develop products that will fit *current* marked needs, but also *future* market needs (For more on Lead-Users see Lilien, Morrison, Searls, Sonnack, & Hippel (2002), Morrison, Roberts, & Midgley (2004))

The nature of ideas in online communities

Several challenges must be overcome before online communities can potentially be harvested for great new ideas. The main challenge is related to the amount of- and the nature of online community data. Online communities often consist of thousands of text pieces that have been exchanged by community members over a time period. These texts are typically organized as *threads* (Lin, Hsieh, & Chuang, 2009). A thread can be started by a community member posting for example a question or a problem. Over time other members of the community answers or comments by responding to the post. This collection of posts, responses and comments is represented as a thread and there can be *many* threads in an online community. It is inside these threads that good ideas are potentially hidden. Facebook posts and comments is one virtual place where one can observe a thread-like information structure.

To find the potentially interesting ideas hiding in online communities, it is relevant to ask the question: What is the *nature* of ideas? Or what does ideas look like when they are written in an online community context? This question is central because the implicit hypothesis of this doctoral project is that ideas written as text in online communities have certain lexical characteristics that make them recognizable for the human eye. And as mentioned earlier, many research papers address how to generate ideas and how to measure idea *quality*, but only few papers comment on the nature of ideas and what separates ideas from other types of information.

A research paper where the characteristics of ideas *are* described is Poetz & Schreier (2012). Here the authors write that ideas contain *need*- and *solution* information. This thought

is very similar to Thorleuchter, den Poel, & Prinzie (2010) who suggests that ideas contain solution information to a defined problem. Need information is a classic focus point for marketing scholars who's focus is often on identifying current and future customer needs so that more- and better products can developed. A customer *need* could be the need for a more powerful computer or for a car with space for six people. By identifying customer related *need* information, the firms *employees* can come up with clever solutions to solving the identified need resulting in better and more profitable products. *Solution* information, on the other hand, is information that solves the need. For example the solution to a need for a more powerful computer, could be to design a computer with a more powerful processor. In a similar manner, a solution to the need for a car with room for six people, could be to design the car with an extra seat in the back.

The literature mentioned above suggests that ideas are formed based on independent information elements. These elements are *needs* and *solutions.* In another study by di Gangi et al. (2010) it is implicitly suggested that ideas can be based on solution information only. The ideas texts shown in Table 1 stem from Dell's own online community. In idea one, an idea is suggested *by* a consumer *to* Dell. The consumer need is the need for a computer with a stable operating system. To appreciate the reason for this need, one would have to know that it is typical for new versions of the Microsoft operating system Windows to contain errors. These make the operating system unstable and therefore a solution to this need it is to wait until Microsoft have fixed all errors in the new operating system, before implementing the system on all new Dell computers. This is a valuable idea for Dell, because Dell learns that there are customers on the market that will favor such a solution. For idea two it also applies that the solution is explicit and the need is implicit. The implicit need is for a laptop computer that can also work as a tablet. The solution is to make the laptop convertible, meaning that the screen

of the laptop can be flipped. If at the same time, the screen is a touch screen, the laptop can function as a tablet.

Table 1 – Idea examples related to Dell computers

Idea one	**Idea two**
´I would like to see both Home and Business computers, especially notebooks, have an XP Home and Pro option on top of Vista until it has at least been out for a year.´	´The XPS and E series notebooks are great, but a move into making 12.1- and 14.1-inch tablet PC convertible notebooks would be fantastic.´

Source: di Gangi et al. (2010) p. 216, Table 1

The notion of *need*- and *solution* information is interesting, because it suggests what the nature of ideas *might* be. It does however also seem premature to define ideas as products of needs and solutions, due to sparse amount of research on the particular topic. An integral part of this work is therefore to investigate what is the nature of ideas in online community texts and add empirical evidence to this discussion.

Aims

This doctoral thesis has four aims. The *first* aim is to investigate if two human raters operating independently from each other can agree on whether an idea is present in a piece of online community text. If that *is* the case, it suggests that ideas have a special lexical pattern and that this pattern *might* be generic for us humans in the sense that we all know it when we see it. The *second* aim is to investigate whether this pattern can be taught to a computer by using proper machine learning and text mining techniques, enabling automatic detection of online community ideas. If this is also the case, it supports the claim that a pattern is present in the text and that there are certain words and expressions in online community ideas that serve as idea predictors. The *third* aim is to investigate what is the nature of idea texts and what is the nature of the lexical patterns that drives automatic detection of ideas? The *fourth* aim is to test if an automated idea detection system detects ideas that will *also* be perceived as

good ideas by firm employees. Data from two different online communities related to beer and Lego is used for the work reported in this thesis.

Machine learning and text mining for automatic idea detection

Machine learning

Machine learning is the process of getting computers (i.e. machines) to recognize patterns in data. When machine learning is coupled with *text mining* or text data, machine learning becomes about teaching computers to find patterns in not just *any* kind of data but *text* data. Machine learning can be seen as the development of a type of artificial intelligence that can carry out complex or simple tasks. The focus in this work is on supervised machine learning and this type of machine learning refers to the case where a *target* variable is present in the dataset. The target variable is used to teach the computer the desired pattern in the texts. An example of this kind of machine learning applied in a real life setting is spam filtering. In a spam filter a machine learning *classifier* has been generated by using a machine learning technique (i.e. algorithm) to separate mails containing spam from emails *not* containing spam. In this way, the classifier can automatically screen *new* incoming e-mails for spam. If a spam mail is detected, it is directed away from the inbox to a separate spam folder or deleted completely. Users of email therefore avoid spending their own time doing this filtration job themselves (See Lai, (2007) for an example of a spam filter study).

Generation of target variable

In a spam filter the target variable has been generated by human raters that have read incoming e-mails and flagged spam e-mails. The result of this is a collection of email texts where two classes of e-mails are present: A spam class and a no-spam class (the non-spam texts is what goes in the inbox). This way of thinking about a text classification problem is interesting because the same principle may be applied to distinguish between idea texts and non-idea text. The difference is that instead of flagging texts for spam vs. non-spam content,

texts should be flagged for idea vs. no-idea content. The result of this exercise is two classes of texts: An *idea text* class and a *non-idea text* class.

The first obstacle for teaching a computer to detect ideas automatically in online communities is to find examples of idea texts and examples of non-idea texts. In the present work idea texts and non-idea texts were detected by using human raters to read *the same* collection of texts and evaluate the texts for idea content. For example Table 2 contains a beer idea text in the left column and a beer non-idea text in the right column. The idea text can be interpreted as a suggestion to a special ingredient in a stout type beer. The non-idea text can be interpreted as a simple comment where a community member states that he/she is *not* going to try a certain beer (Fosters is s special type of beer brand). In Table 3 a Lego idea text is displayed in the left column and a Lego non-idea text is displayed in the right column. One can interpret the idea text as a wish from a community member, who states Lego should bring back a discontinued product. The Lego non-idea text can be interpreted as dialogue between community members discussing where to find cheap Lego bricks.

Table 2 – Beer texts with idea content and no idea content

Beer idea text	**Beer non-idea text**
´I don't know about pineapple, But I have used unsweetened bakers chocolate powder in a stout that turned out pretty well.´	´I can verify that. I'm an Aussie and never had a fosters. It's not because I don't want to at least try it, but no pubs have it on tap! Not about to waist money buying a carton. Pure marketing.....´

Table 3 – Lego texts with idea content and no idea content

Lego idea text	**Lego non-idea text**
´Dear Lego, If you're bringing back a Technic set, forget about sets that were in the shops two years ago. Bring back the 8868 Airtech Claw Rig or the 8480 Space Shuttle.´	´Wow that funny, I was just searching Bricklink last night for these very same parts...they are there, but they're not cheap. Yeah I'd be all for bulk packs of them from´

If two raters are used for idea rating on many texts, the result of the exercise described above is two piles of texts for each rater assigned to the idea detection task. An idea text pile and a non-idea text pile. Based on a comparison of these four piles, the *agreement* of the raters can be assessed. In the example in Table 4 we see that a rater number one, has detected 44 idea texts (C7) and 56 non-idea texts (C8). A rater number two has detected 36 idea texts (C3) and 64 non-idea texts (C6). They agreed on 28 idea texts (C1) and 48 non-idea texts (C5), but disagreed on 24 texts (C2 and C4). This corresponds to an agreement equal to 0.76.

Table 4 – Rater agreement table based on imaginary example. Agreement is 0.76

		Rater one		
		Idea	Non-idea	∑
Rater two	Idea	28 (C1)	8 (C2)	36 (C3)
	Non-idea	16 (C4)	48 (C5)	64 (C6)
	∑	44 (C7)	56 (C8)	100 (C9)

The question that remains is if 0.76 agreement is acceptable agreement. In the example displayed in Table 4 it might be acceptable, but what if the raters behaved like the examples showed in Table 5 and Table 6. In the example showed in Table 5 the raters obtain 96% agreement, but they only agree on one idea text out of the five idea texts they detected together. In Table 6 two raters obtain 50% agreement, but they also have 50% *dis*agreement. In this case it may look as if the two raters have scattered the ratings at random.

Table 5 – Rater agreement table based on another imaginary example. Agreement = 0.96

		Rater one		
		Idea	Non-idea	∑
Rater two	Idea	1	2	3
	Non-idea	2	95	97
	∑	3	97	100

Table 6 – Rater agreement table based on imaginary example. Agreement = 0.50

		Rater one		
		Idea	Non-idea	∑
Rater two	Idea	25	25	50
	Non-idea	25	25	50
	∑	50	50	100

Cohens kappa (κ) is an inter-rater agreement measure that can take into account chance agreement and it provides a framework for assessing agreement between two independent raters in a non-subject manner (J. Cohen, 1960; J. Cohen, 1968; von Eye & von Eye, 2008). It has been has suggested that: $\kappa < 0$ is poor agreement, $0 < \kappa \leq 0.20$ is slight, $0.20 < \kappa \leq 0.40$ is fair, $0.40 < \kappa \leq 0.60$ is moderate, $0.60 < \kappa \leq 0.80$ is substantial, $0.80 < \kappa \leq 1$ is almost perfect agreement (Landis & Koch, 1977). If κ values are calculated for the examples described above $\kappa = 0.50$ (Table 4), $\kappa = 0.31$ (Table 5) and $\kappa = 0$ (Table 6).

A second aspect related to κ is that it has a theoretical maximum ($\kappa_{max)}$ that depends on the marginal distributions of the ratings. To take this into account it has been suggested that both regular κ as well as κ as a proportion of maximum possible (κ/κ_{max}) is reported. In the examples above it is only the example in Table 4 where κ_{max} is below 1. In that particular case, $\kappa_{max} = 0.83$ so that $\kappa/\kappa_{max} = 0.60$.

Text pre-processing

Before the texts can be used for machine learning they have to be pre-processed. In a spam filter the classifier learns what are the words and expressions associated with spam emails (We call words and expressions *terms* from now on). For example one can imagine that “earn money”, “subscribe to“ or “win a prize” are typical terms that may occur in a spam e-mails. However before machine learning can be applied, the raw text will have to be pre-processed so the text can be used for machine learning. When the raw text is pre-processed, the collection of texts are turned into a row / column format, where all unique terms are represented as columns and all unique texts are listed as rows. All punctuation marks, numbers and all extra whitespaces are removed. Upper case letters are transformed to lower case letters or vice-versa . N-grams can also be generated which refers to series of words. For example “beer” is a one-gram, “good beer” is a two-gram and “good beer idea” is a three-gram. N-grams are useful because they carry additional *meaning* that each a single term do not (Zanasi, 2007; Feinerer, Hornik, & Meyer, 2008). In the present work all text mining operations were performed via the tm package in R (Feinerer & Hornik, 2015).

Partitioning, training, tuning and testing

Partitioning refers to the small but important task of separating the full text dataset into separate independent datasets (See Figure 3 for overview). This step is central for proper tuning (also called calibration) as well as assessing the performance of the trained classifiers. Machine learning algorithms can become so complex that they can fit the data training data *perfectly*. This will result in perfect performance if the classifier was used on texts that look exactly like the texts it was trained on. This phenomenon is called *over-fitting* and it is not preferred because in practice a classifier will most often *not* be applied on texts exactly like the ones it was trained on but *new* texts only *similar* to the texts it was trained on (Hastie, Tibshirani, & Friedman, 2008).

To avoid over-fit the full dataset can be partitioned into preferably three independent sets. The first set is a training set. The second a validation set. When a third set is present, this is called a hold-out set or a test set (In this thesis the term *hold-out* set is used). The training set is the input for the machine learning technique. The validation set is used to tune the classifier so optimal performance is obtained on the validation set rather than on the training set. *Tuning* refers to the process of adjusting one or several parameters that are often individual for the applied classification technique. Some classifiers have several tuning parameters and the optimal combination has to be determined. A grid search can be used for this purpose, meaning that a classifier is trained for all possible combinations of tuning parameters. Each classifier is tested on the validation set and the classifier that obtains highest performance is considered the best classifier. Over-fit and *true* performance is assessed by applying the best classifier on the hold-out set (Linoff & Berry, 2011).

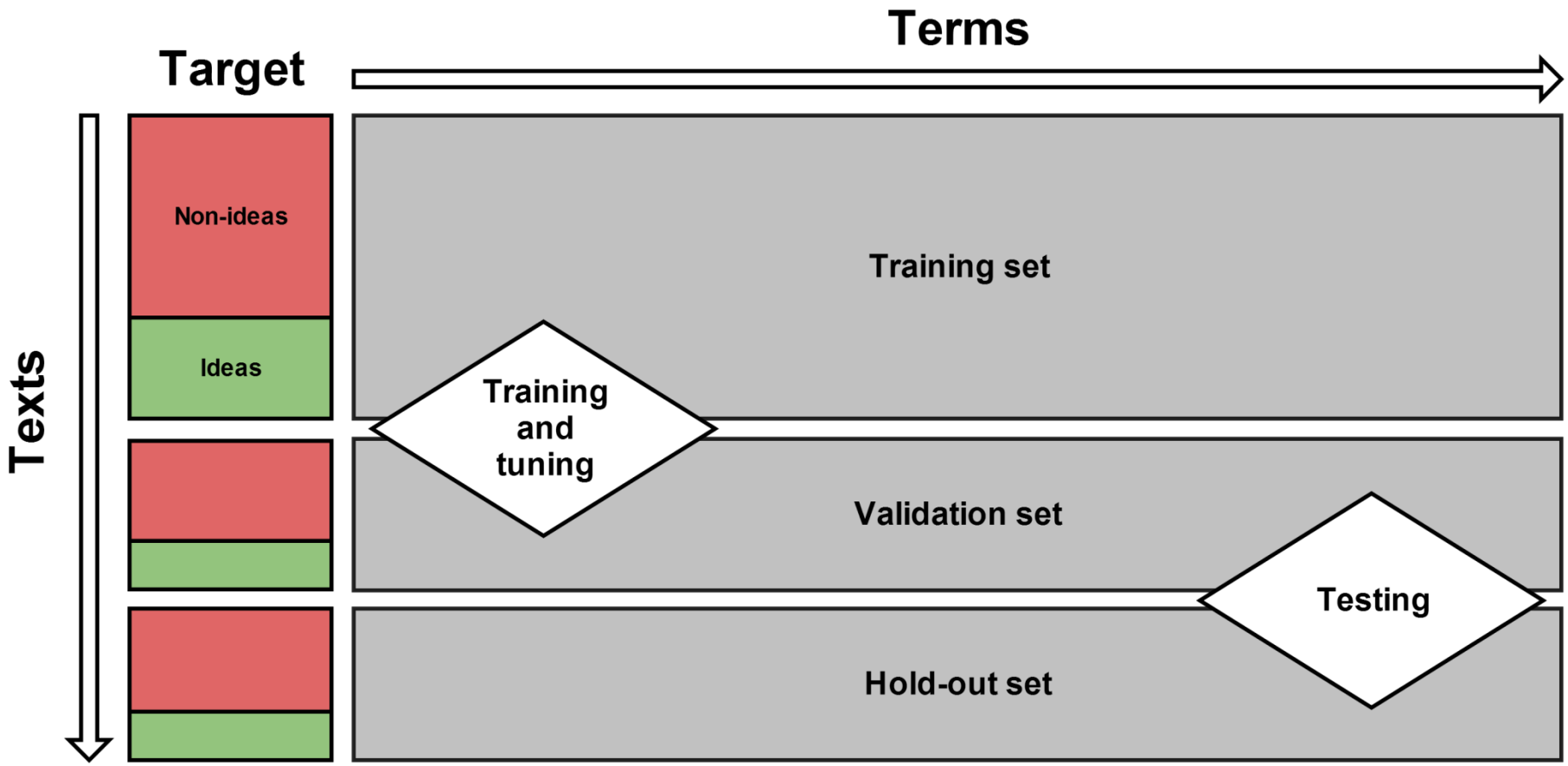


Figure 3 – The figure illustrates the principles of partitioning, training, tuning and testing. The grey boxes illustrate the pre-processed text dataset where terms have been counted. Columns (horizontal axis) represent target and terms. Rows (vertical axis) represent texts. The target column illustrates the relative amount of idea texts and non-idea texts within each partition. The natural balance in the target variable is the same for each partition. The parallelogram "training and tuning" illustrates that the training set and validation set are used for training and tuning. The parallelogram "testing" illustrates that the validation set and the hold-out set are used for testing and assessing over-fit.

Choice of classification technique

Several machine learning techniques can perform text classification tasks. They all vary in their nature, their tendency to over-fit, and the number of adjustable tuning parameters. Support Vector Machines, Partial Least Squares, Decision Trees, Neural Networks and Naïve Bayes are all types of machine learning algorithms that can be used for classification of idea texts. In the present work the focus was on two techniques in particular. The one technique is called *Support Vector Machines* (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995) and the second is called *Partial Least Squares* (Wold, Martens, & Wold, 1983; Wold, Sjöström, & Eriksson, 2001)

Partial least squares, also called PLS, has not been used for text classification in the same degree as has support vector machines. And it seems like partial least squares has been overlooked by the machine learning and text mining community. See for example Witten & Frank (2005), Feldman & Sanger (2006), Han & Kamber (2006) Kao & Poteet (2007) and Linoff & Berry (2011). These are all examples textbooks on data mining, text mining and/or machine learning that do *not* take into account the partial least squares technique. Only Hastie et al. (2008) describes the technique.

Support Vector Machines

Support vector machines are known for their high performance on high dimensional sparse datasets, and are therefore a good choice for text classification tasks (See Table 2 in Paper 1). Support vector machines were developed in the 90's and they come in linear and non-linear varieties. The linear support vector machine is the least complex variety and when used for text classification it takes term vectors as input together with the related target value (i.e. idea or non-idea). The term vectors represent the raw term counts for each text in the dataset. Support vector machines are comparable to linear discriminant analysis, *but* they allow for data that cannot be perfectly separated.

A linear support vector machine calculates the distance between the texts by a dot product calculation based on the term vectors. The bigger the distance in this space the more different the texts are. When the dot product has been calculated, the linear support vector machine finds the line surrounded by a margin that best separates the idea texts from the non-idea texts. The width of the margin is determined by a cost parameter *C*, which is a tuning parameter that requires optimization. Optimization can be done by specifying a series of C values and for each C value a classifier is trained and used for classification on the validation set. The C that obtains the highest performance on the validation set is the optimal C. Performance is then tested on the hold-out set (See Ben-Hur & Weston (2010) for more on support vector machines). In the present work the e1071 R-package was used for implementing support vector machines (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2015).

Partial Least Squares

Partial least squares stems from chemometrics and sensometrics. It is able to handle many correlated predictor variables (i.e. terms) and few observations (i.e. texts) and it can be used for both regression and classification. Partial least squares reduce the original set of terms to a smaller set of *latent* variables or components. The components are constructed so that the covariance between the terms and the target variable is maximized. The result of this computational procedure is a loadings matrix and a scores matrix. In the loadings matrix, the original terms are represented row-wise and the components are represented column-wise. The loadings matrix represents the magnitude of each individual term in relation to each individual component. This matrix allows for insights about which terms drives the latent phenomenon embedded in each individual component. In the scores matrix the components are displayed in columns, but instead of the terms, it is the *texts* that are displayed row-wise. Thus, in this particular matrix, the relation between texts and components allows for insights

about the latent phenomenon embedded in each text. The number of components is a tuning parameter where the optimal number can be identified computationally by a grid-search. For example one can train a classifier based on one component and another classifier based on two components etc. Each classifier is used for classification on the validation set. The number of components that obtains the highest performance on the validation set is the best classifier. Performance is tested on the hold-out set.

Another central argument for choosing partial least squares is that several variable selection procedures have been developed for the method. In a text classification context this is relevant since terms are often many. In the present work all datasets contained over 9.000 terms. This amount of variables/terms is considered high and interpretation becomes a challenge if the set of terms is not reduced to include only the most predictive terms. Several procedures exists that can be used in conjunction with partial least squares and (Mehmood, Liland, Snipen, & Sæbø, 2012). In the present work *Significance Multivariate Correlation* was used (Tran, Afanador, Buydens, & Blanchet, 2014). The pls R package was used for implementing partial least squares (Mevik, Wehrens, & Liland, 2015)

Class imbalance and cut-off threshold

Support vector machines and partial least squares yield real numbers as output (i.e. 0.25, 0.5, 0.776 etc.) and not binary classifications (i.e. idea vs. non-idea). Therefore it is up to the user of the method to set a proper cut-off threshold that determines what texts will belong to the idea text class and what texts will belong to the non-idea text class. In the present work the cut-off threshold was treated as a tuning parameter similar to cost for support vector machines and the number of components for partial least squares.

By adjusting the cut-off threshold a well-known problem related to classification problems can be handled. This problem is known as the class imbalance problem and it refers

to a problem that occurs when the distribution in the target class is skewed (i.e. there are more non-idea texts than idea texts or vice versa). As a consequence of this skewness, the classification techniques tend to favour correct classification of the *majority* class over correct classification of the minority class (Tian, Gu, & Liu, 2010; Diao, Yang, & Wang, 2012; Menardi & Torelli, 2014). For example in the case where 990 non-idea texts and 10 idea texts are present in a data set, a classifier may obtain 99% accuracy by classifying all texts as non-idea texts. This is not preferable and several strategies have been suggested for handling this problem. One strategy is to adjust the cut-off threshold and tune the classifier for maximum performance on a performance measure that favours the minority class (Performance measures will be explained in the next section) (Provost, 2000). Another strategy is to use an *under-bagging* strategy as described by Galar, Fernandez, Barrenechea, Bustince, & Herrera (2012). In the present work both approaches were used. In Paper 1 an under-bagging approach was used. In Paper 3 and Paper 4 cut-off threshold adjustment was used.

Performance measures

The performance measures used for measuring classification performance stems from signal detection theory (Stanislaw & Todorov, 1999). In the contexts of this thesis, the signal is an *idea signal* in an online community text, and the *non-idea* signal is everything else (i.e. noise). Performance measures are obtained by comparing the classifications of a classifier with the *true* class the texts belong to. If a classifier has detected an idea text in the hold-out set *and* the human raters have *also* detected the same idea texts, the text can be considered a true positive (TP). Correctly detected *non*-idea texts are called true negatives (TN). If a classifier detected an idea text that was *not* detected as an idea text by the human raters, the text can be considered a false positive (FP). In the case where a classifier has detected a non-idea text that was detected as an idea text by the human raters, the text can be considered a false negative (FN). The total counts of TP's, TN's, FP's and FN's can be represented as a

confusion matrix. The confusion matrix is identical to the agreement tables used for the κ calculations already presented.

In the example displayed in Table 7, two human raters have read the same 100 texts and detected the same 35 idea texts and the same 65 non-idea texts (perfect agreement). A machine learning classifier that has been trained to detect idea texts, has also "read" the same 100 texts and detected 40 idea texts and 60 non-idea texts. The raters and the classifier agree on 25 idea texts (TP) and 50 non-idea (TN) texts. They disagree on 25 (FN and FP) texts. Based on the information in the confusion matrix, performance measures can be calculated. In this work the focus was on six performance measures. These are *accuracy*, *recall*, *precision*, F_1 as well as κ and κ/κ_{max}.

Table 7 – Example of confusion matrix that is used to compare classification of classifier and raters

Confusion matrix		Classifier		
		Idea	Non-idea	∑
Raters	Idea	25 (TP)	10 (FN)	35
	Non-idea	15 (TP)	50 (TN)	65
	∑	40	60	100

The interpretation of accuracy (Equation 1) is how many texts the classifier classifies correctly. It is a measure that if very often used, *but* it can be misleading if the idea text and non-idea text distribution is skewed. For example in the case where 5 idea texts and 95 non-idea texts are evaluated, the classifier obtains accuracy = 0.95, by classifying all texts as non-idea texts. Accuracy = 0.95 can be considered high, but the classifier is useless because it does not manage to find the idea texts one is interested in. This is why it can be useful to calculate other performance measure in *relation to* the idea class and use a such measure for

optimization. When recall is calculated with respect to the idea text class, it can be interpreted as how big a part of the true idea texts the classifier managed to identify (Equation 2). Precision in relation to the idea text class can be interpreted as how big a part of the detected idea texts by the classifier are true idea texts (Equation 3). F_1 in relation to the idea class can be interpreted as a measure that balances precision and recall (Equation 4). All classifiers described in the present work were optimized with respect to the F_1 measure. The argument for this choice is related to the class imbalance problem described in the previous section.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$F_1 = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \quad (4)$$

Summary of results

The *first aim* of this doctoral project was to investigate if two human raters can agree on whether an idea is present in a piece of text stemming from an online community. The results in Paper 1 and in Paper 4 suggest that this is largely the case. In Paper 1, two human raters were asked to read the same 3,000 Lego texts. The raters obtained $\kappa = 0.55$ and $\kappa/\kappa_{max} = 0.63$. According to Landis & Koch (1977) this can be considered moderate to substantial agreement. In Paper 4, two brewing experts were asked to read the same 200 beer- and brewing texts. They obtained $\kappa = 0.37$ and $\kappa/\kappa_{max} = 0.74$. This can be considered fair and substantial agreement.

The *second aim* was to investigate whether the lexical pattern embedded in idea texts can be taught to a computer by using machine learning and text mining techniques. The results presented in Paper 1 and Paper 4 are appropriate for addressing this aim. The results reported in Paper 1, suggest that machine learning can to an acceptable extent be used for detecting ideas (Table 8). In this study $\kappa = 0.56$ and $\kappa/\kappa_{max} = 0.76$ on the hold-out set. These results suggest that there was moderate to substantial agreement between the raters and the classifier. The classifier trained in Paper 4 obtained $\kappa = 0.34$ and $\kappa/\kappa_{max} = 0.73$ for expert one and $\kappa = 0.48$ and $\kappa/\kappa_{max} = 0.51$ for expert two (Table 9). This corresponds to fair and substantial agreement.

Table 8 – Classifier performance related to Paper 1

Set	TP	TN	FP	FN	F_1	Accuracy	Recall	Precision	κ	κ/κ_{max}
Validation	27	628	38	7	0.55	0.94	0.79	0.42	0.51	0.77
Hold-out	228	162	22	88	0.81	0.78	0.72	0.91	0.56	0.76

Table 9 – Classifier performance related to Paper 4

Set	TP	TN	FP	FN	F_1	Accuracy	Recall	Precision	κ	κ/κ_{max}
Validation	56	143	5	4	0.93	0.96	0.93	0.92	0.90	0.91
Hold-out (Expert one)	35	102	57	6	0.53	0.69	0.85	0.38	0.34	0.73
Hold-out (Expert two)	64	85	28	23	0.72	0.75	0.74	0.70	0.48	0.51

The *third aim* was to investigate what is the nature of the syntactical pattern contained in idea texts in the two online communities used in this work (i.e. beer and Lego). In relation to this aim, the interpretation of the results in Paper 3 suggests that idea texts contain terms that reflects suggestion and solution dialogue. These results were derived by assessing a *subset* of the texts for each online community where short- and long texts were removed. For the beer community idea texts, “if you”, “solution”, “you want”, “you can” and “thinking” are terms that reflects suggestion and solution dialogue for how to do something in the brewing process. For the Lego community, suggestion solution terms are for example “would be”, “they would”, “i think”, “idea” and “could be”. Here the nature of idea texts seems to be related to new product wishes from Lego users to the Lego firm.

Last, the *fourth aim* was to test if employees working for a firm are likely to perceive the ideas detected as also being *good* ideas. The results presented in Paper 4 suggest that the ideas detected in this study were not considered particularly novel. The ideas were considered rather feasible and have medium value. The detected ideas were rated as having a medium *overall* quality.

Discussion and future perspectives

The first aim of the present work was to investigate if two human raters will find the same idea texts and the same non-idea texts. If this *is* the case, it suggests that the concept of an idea in an online community text might to some extent be generic to the human mind. And it can be expected to use machine learning and text mining to generate classifiers that can detect ideas that will be perceived as ideas by people who never took part in training of the machine learning classifier. In the present work, this principle was demonstrated on two online communities related to the product domains of beer and Lego. The results suggest that human raters do to an acceptable extent recognize the same idea texts and non-idea texts. This is supported by the kappa interrater agreement measures reported in Paper 1 as well as Paper 4. Future research may ask the question if this is also the case for online communities where the topical nature is different from beer and Lego? What about ideas related to other topics like wine, food or smartphones? Or what about ideas related to more abstract topics like management, innovation or sustainability?

A central premise for automatically detecting ideas in online communities, is that the potential lexical pattern contained in idea texts and non-idea texts, can be taught to a machine learning technique and by that, generate a machine learning classifier. The results presented in Paper 1 and Paper 4 indicate that this is the case to an acceptable extent. In Paper 1, a linear support vector machine classifier performed $F_1 = 0.81$ on an external hold-out set. The related kappa measures are $\kappa/\kappa_{max} = 0.77$ for the internal validation set and $\kappa/\kappa_{max} = 0.76$ on the external hold-out set. In Paper 4, $F_1 = 0.53$ for expert one and $F_1 = 0.72$ for expert two. These results can be compared to Thorleuchter & Van den Poel (2013) who investigated if text mining could be applied for identifying ideas that fits a specific problem description. F_1 measures in the range from 0.29 to 0.38 are reported. In another text mining study, *not* related

to ideas but online chat in general, kappa measures in the range $\kappa = 0.53$ to $\kappa = 0.62$ are reported (Tirunillai & Tellis, 2014).

If the method presented in this work is to develop towards obtaining higher accuracy and/or agreement, a clearer definition of ideas embedded in texts must be developed. The raters in Paper 1 disagreed on 197 texts and these texts were omitted from further analysis. In Paper 3, over 50% of the training texts were omitted because the raters could not agree on the class of the texts. Neither could the experts in Paper 4 agree on the class of the 200 texts. It seems plausible that the disagreement is related to the ambiguity of the idea text concept and it generates a challenge for future research. Future research should agree on a definition of online community idea texts and ask the question: What exactly is an idea in an online community text? Poetz & Schreier (2012) and van den Ende et al. (2015) writes that ideas are related to problems and solutions. This is in line with creativity literature as for example Wallas (1926) and Lubart (2001), who describe the creative process a the process from problem (start) to solution (end). Another viewpoint on this matter is that ideas *must* signal some degree of novelty and usefulness (Aldous, 2007). In summary the five sources suggest that ideas must contain a problem description as well as a solution to the problem. And in addition the solution must be novel and useful. No such criteria were used in this work. It remains for future research, to develop a proper definition of ideas written in online community texts. This may be used for future studies of similar nature as the studies presented in this work.

The results presented in Paper 3 may be used for defining online community ideas. Here it is suggested that idea texts often contain suggestion/solution terms as well as product domain related terms. Another observation highlighted in Paper 3 is that many of the idea texts in the Lego community, seems to be related to new products whereas many of the ideas in the brewing community is related to the brewing *process*. Future research should

incorporate the aspect of idea category into the generation of training datasets so that idea raters *in addition* to rating idea texts and non-idea texts, should also determine the category of the idea texts. Categories such as *product idea*, *process idea*, *service idea* and *pricing model idea* may be suitable categories.

An aspect that could be taken into consideration when defining online community ideas is the nature of the *non*-idea texts. The results presented in Paper 3 indicate that ideas texts may be characterized by suggestion/solution dialogue. But what kind of information do the non-idea texts contain? Does the non-idea text class contain needs, complaints, problems, facts or opinions? What exactly is the nature of the content in the non-idea class? Preliminary analysis not reported in this work indicates that the beer community non-idea texts may to a large extent contain question making. For example, members of the brewing community often write about: "How can I do something…?", "Why is this happening to my beer…?" or "Does anyone have a good suggestion for…?". The same does however not seem to apply for the non-idea texts in the Lego community, and no preliminary results have given any indication on nature of the *non-ideas* related to the Lego community.

The fourth and final aim pursued in this work was to investigate, if the ideas that are potentially detected by the developed method will also be perceived as *good* ideas by firm employees? In Paper 4 a good idea was defined as an idea that is perceived as novel, feasible and valuable by appropriate product experts. The results presented in Paper 4 suggest that the method may detect ideas that are perceived as novel, feasible and/or valuable by product experts. Future research should aim at validating these results. Also should future research investigate what is the most appropriate idea quality attributes for measuring the quality for fuzzy front end ideas from online communities. The work carried out by Kristensson et al. (2004), Björk & Magnusson (2009), di Gangi et al. (2010), Poetz & Schreier (2012), Magnusson, Netz, & Wästlund (2014), Magnusson, Wästlund, & Netz (2014), Frederiksen &

Knudsen (2017) and in particular the study conducted by Dean et al. (2006) may be useful for developing such a measurement tool.

If future research can address the issues outlined above, new doors may be opened for researchers and practitioners. Similar methods may be used to automatically detect ideas hiding in the worldwide digital ecosystem of online communities. The potential positive consequences of developing and employing such a technology, is that it may increase firms ability to react to new market needs and changes in the external environment. Something that is critical for business success (W. M. Cohen & Levinthal, 1990; Roberts, Galluch, Dinger, & Grover, 2012). From a societal perspective this is a highly favorable situation because firms do not waste resources on developing products that will not be accepted by the end-user. On the other hand side, a concern related to further development of this type of method, is that the method can be seen as an attempt to develop a type of artificial intelligence that can *steal* ideas. If this shows to be a valid concern, it may lead to a *decrease* in knowledge sharing and idea generation in online communities. For now however this concern remains a hypothesis that can- and should be tested. In relation to this, a natural follow up question on the present research is: “What are the potential consequences of employing artificial intelligence systems for automatic idea detection?”

Conclusion

Ideas are the starting point for innovation and the starting point for this doctoral project was to investigate if it is possible to teach a computer to automatically detect ideas written as text in online communities. If ideas can be automatically detected, methods can be developed, that can aid firms in detecting ideas in online conversation and hopefully accelerate innovation for the common good of society. In the present work this was done by developing and employing an artificial intelligence system based on text mining and machine learning that has learned to detect ideas related to beer and Lego.

The presented results indicate that machine learning and text mining can play an important role when it comes to identifying new and interesting ideas useful for innovation. The applied method detected Lego related ideas as well as beer related ideas. It was also demonstrated that the method may identify beer ideas that are perceived as relatively novel, feasible and valuable by firm employees. The results also indicate that the method will fail to identify some ideas as well as it will select texts that are *not* ideas.

By employing the method firms can gain access to excessive amount of ideas useful for innovation and reduce the manual labor costs that would elsewise be associated with identifying the ideas, should firm employees have read the texts manually. The proposed method opens several doors that can be investigated. Future research should test and seek to improve the accuracy of the method on other product domains than beer and Lego. Future research should also develop a proper definition of online community ideas and attempt to develop measurement scales that can be used to measure the quality of fuzzy front end ideas. Finally, future research should be open to the claim that the development of artificial intelligence for automatic idea detection, may result in a decrease in online knowledge sharing.

References

Albors, J., Ramos, J. C., & Hervas, J. L. (2008). New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, *28*(3), 194–202.

Aldous, C. R. (2007). Creativity, problem solving and innovative science: Insights from history, cognitive psychology and neuroscience. *International Education Journal*, *8*(2), 176–187.

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.

Baregheh, A., Rowley, J., & Sambrook, S. (2009). Towards a multidisciplinary definition of innovation. *Management Decision*, *47*(8), 1323–1339.

Bayus, B. L. (2013). Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community. *Management Science*, *59*(1), 226–244.

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. *Methods in Molecular Biology*, *609*, 223–239.

Björk, J., & Magnusson, M. (2009). Where Do Good Innovation Ideas Come From? Exploring the Influence of Network Connectivity on Innovation Idea Quality. *Journal of Product Innovation Management*, *26*(6), 662–670.

Boeddrich, H.-J. (2004). Ideas in the Workplace: A New Approach Towards Organizing the Fuzzy Front End of the Innovation Process. *Creativity and Innovation Management*, *13*(4), 274–285.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory*.

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

Cohen, J. (1968). WEIGHTED KAPPA: NOMINAL SCALE AGREEMENT WITH PROVISION FOR SCALED DISAGREEMENT OR PARTIAL CREDIT. *Psychological Bulletin*, *70*(4), 213.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, *35*(1), 128.

Cooper, R. G. (2008). Perspective: The Stage-Gate® Idea-to-Launch Process—Update, What's New, and NexGen Systems*. *Journal of Product Innovation Management*, *25*(3), 213–232.

Cooper, R. G., & Kleinschmidt, E. J. (1986). An Investigation into the New Product Process: Steps, Deficiencies, and Impact. *Journal of Product Innovation Management*, *3*(2), 71–85.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297.

Dahlander, L., Frederiksen, L., & Rullani, F. (2008). Online Communities and Open Innovation. *Industry & Innovation*, *15*(2), 115–123.

Dahlander, L., & Magnusson, M. (2008). How do Firms Make Use of Open Source Communities? *Long Range Planning*, *41*(6), 629–649.

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.

Diao, L., Yang, C., & Wang, H. (2012). Training SVM email classifiers using very large imbalanced dataset. *Journal of Experimental & Theoretical Artificial Intelligence*, *24*(2), 193–210.

Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package (Version 0.6-2) [R]. Retrieved from https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Feldman, R., & Sanger, J. (2006). *THE TEXT MINING HANDBOOK: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

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

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.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2. edition). San Francisco, CA: Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning - Data Mining, Inference and Prediction* (Second edition). Stanford, CA: Springer.

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.

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. London, UK: Springer-Verlag.

Kim, J., & Wilemon, D. (2002). Focusing the Fuzzy Front-End in New Product Development. *R&D Management*, *32*(4), 269–279.

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

Lai, C.-C. (2007). An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems*, *20*(3), 249–254.

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

Lee, G. K., & Cole, R. E. (2003). From a Firm-Based to a Community-Based Model of Knowledge Creation: The Case of the Linux Kernel Development. *Organization Science*, *14*(6), 633–649.

Lilien, G. L., Morrison, P. D., Searls, K., Sonnack, M., & Hippel, E. von. (2002). Performance Assessment of the Lead User Idea-Generation Process for New Product Development. *Management Science*, *48*(8), 1042–1059.

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.

Linoff, G., & Berry, M. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3. Edition). Indianapolis, IN: Wiley publishing.

Lubart, T. I. (2001). Models of the Creative Process: Past, Present and Future. *Creativity Research Journal*, *13*(3–4), 295–308.

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., Netz, J., & Wästlund, E. (2014). Exploring holistic intuitive idea screening in the light of formal criteria. *Technovation*, *34*(5–6), 315–326.

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.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2015). pls: Partial Least Squares and Principal Component Regression (Version 2.5-0) [R]. Retrieved from https://CRAN.R-project.org/package=pls

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071) (Version 1.6-7) [R]. Retrieved from https://CRAN.R-project.org/package=e107

Moenaert, R. K., De Meyer, A., Souder, W. E., & Deschoolmeester, D. (1995). R&D/Marketing Communication During the Fuzzy Front-End. *Engineering Management, IEEE Transactions on*, *42*(3), 243–258.

Morrison, P. D., Roberts, J. H., & Midgley, D. F. (2004). The nature of lead users and measurement of leading edge status. *Research Policy*, *33*(2), 351–362.

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*.

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.

Provost, F. (2000). Machine Learning from Imbalanced Data Sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*.

R Core Team. (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Reid, S. E., & De Brentani, U. (2004). The Fuzzy Front End of New Product Development for Discontinuous Innovations: A Theoretical Model. *Journal of Product Innovation Management, 21*(3), 170–184.

Ridings, C. M., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *The Journal of Strategic Information Systems, 11*(3), 271–295.

Roberts, N., Galluch, P. S., Dinger, M., & Grover, V. (2012). Absorptive Capacity and Information Systems Research: Review, Synthesis, and Directions for Future Research. *MIS Quarterly*, *36*(2), 625–648.

Schumpeter, J. A. (1943). *Captitalism, Socialism & Democracy*. USA: Routledge, London and New York.

Smith, P. G., & Reinertsen, D. G. (1991). *Developing Products in Half the Time: New Rules, New Tools, 2nd Edition*. New York: Van Nostrand Reinhold.

Soukhoroukova, A., Spann, M., & Skiera, B. (2012). Sourcing, Filtering, and Evaluating New Product Ideas: An Empirical Exploration of the Performance of Idea Markets. *Journal of Product Innovation Management*, *29*(1), 100–112.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Thorleuchter, D., den Poel, D. V., & Prinzie, A. (2010). Mining ideas from textual information. *Expert Systems with Applications*, *37*(10), 7182–7188.

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

Tian, J., Gu, H., & Liu, W. (2010). Imbalanced classification using support vector machine ensemble. *Neural Computing and Applications*, *20*(2), 203–209.

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Tran, T. N., Afanador, N. L., Buydens, L. M. C., & Blanchet, L. (2014). Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemometrics and Intelligent Laboratory Systems*, *138*, 153–160.

Van de Ven, A. (1986). CENTRAL PROBLEMS IN THE MANGEMENT 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.

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

von Hippel, E. (1986). Lead Users: A Source of Novel Product Concepts. *Management Science*, *32*, 791–805.

von Hippel, E., & Foster, R. N. (1988). The sources of innovation. *McKinsey Quarterly*, (1), 72–79.

von Hippel, E., & von Krogh, G. (2003). Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science. *Organization Science*, *14*(2), 209–223.

Wallas, G. (1926). *The art of THOUGHT*. England: Solis Press.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2. edition). San Francisco, CA: Morgan Kaufmann publishers.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In B. Kågström & A. Ruhe (Eds.), *Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22–24, 1982* (pp. 286–293). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130.

Zanasi, A. (2007). *Text Mining and its Applications to Intelligence, CRM and Knowledge Management* (1. edition). Southampton, UK: WIT Press.

Paper 1

Kasper Christensen, Sladjana Nørskov, Lars Frederiksen and Joachim Scholderer.

In Search of New Product Idea: Identifying Ideas in Online Communities by Machine learning and Text Mining (2016). Creativity and Innovation Management (Available Online)

The paper is reprinted with permission from John Wiley and Sons

In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining

Kasper Christensen, Sladjana Nørskov, Lars Frederiksen and Joachim Scholderer

Online communities are attractive sources of ideas relevant for new product development and innovation. However, making sense of the 'big data' in these communities is a complex analytical task. A systematic way of dealing with these data is needed to exploit their potential for boosting companies' innovation performance. We propose a method for analysing online community data with a special focus on identifying ideas. We employ a research design where two human raters classified 3,000 texts extracted from an online community, according to whether the text contained an idea. Among the 3,000, 137 idea texts and 2,666 non-idea texts were identified. The human raters could not agree on the remaining 197 texts. These texts were omitted from the analysis. The remaining 2,803 texts were processed by using text mining techniques and used to train a classification model. We describe how to tune the model and which text mining steps to perform. We conclude that machine learning and text mining can be useful for detecting ideas in online communities. The method can help researchers and firms identify ideas hidden in large amounts of texts. Also, it is interesting in its own right that machine learning can be used to detect ideas.

Introduction

Ideas are the seeds of innovation and an important determinant of success in managing innovation. Previous research shows that access to a continuous flow of new product ideas can help companies reduce R&D costs and develop more attractive products for their customers (Sawhney & Prandelli, 2000; di Gangi, Wasko & Hooker, 2010). Online communities are a particularly interesting source of ideas because companies can gain 'access to new information, expertise, and ideas not available locally' within their own organizational boundaries (Wasko & Faraj, 2005, p. 36; see also Jeppesen & Frederiksen, 2006). Although online communities can also exist within organizational boundaries, for example as platforms for team collaboration (Björk & Magnusson, 2009), a more prominent case in the innovation literature are the open collaborative development processes in open source software communities (e.g., Hertel, Niedner & Herrmann, 2003; von Krogh, Spaeth & Lakhani, 2003; Henkel, 2006; Foss, Frederiksen & Rullani, 2015). Similarly, in the marketing literature, online communities built around particular brands – e.g. Audi or Lego (Füller et al., 2006; Antorini, Muñiz & Askildsen, 2012) – are considered important sources of ideas for brand or line extensions and product re-positioning (Ogawa & Piller, 2006).

Although these ideas are technically freely available in online communities, two challenges must be overcome. These challenges stem from the nature of online community data. First, the often excessive amounts of information exchanged in online communities can make it difficult to identify which pieces of information are actually relevant (Dahlander & Piezunka, 2015). Thus, it is often the time required for sifting the available information, rather than gathering the information as such, that drives the costs of incorporating it in the innovation process. Second, information in online communities tends to be in the form of unstructured

© 2016 John Wiley & Sons Ltd

10.1111/caim.12202

text and requires substantial pre-processing before it can be statistically analysed (Netzer et al., 2012). The traditional way of meeting these challenges would be to code the information manually. However, manual coding of unstructured text into structured data is at best expensive and at worst infeasible as online communities may consist of several thousand members who exchange millions of messages and comments.

The aim of the research presented here is to introduce a method that can perform these tasks *automatically*. Based on a relatively small set of manually coded training data, a classification algorithm is developed that distinguishes texts that include ideas from texts that do not include ideas. The algorithm can then be used on arbitrarily large collections of text to identify those texts likely to include ideas, substantially reducing the time that would have been required if all texts had been coded manually.

The next section will review previous applications of text mining in the innovation literature and discuss the particular target event our algorithm is intended to detect: the presence of an idea. In the method section, we provide a non-technical introduction to the text mining and machine learning techniques on which the algorithm is based and describe the training and tuning process. We then report the training results and the performance of the trained classifier in an independent test set. Finally, we discuss our findings and offer concluding remarks.

Previous Applications of Text Mining in Innovation

Only recently has the innovation literature adopted text mining techniques. Antons, Kleer and Salge (2016) describe the topic structure of papers published in the *Journal of Product Innovation Management*, using latent Dirichlet allocation. The latent Dirichlet allocation can be understood as a discrete analogue to principal component analysis that allows for automatic mapping of the topic structure in a collection of texts. Tirunillai and Tellis (2014) used the same technique to describe the topic structure in a collection of online product reviews. Also, Kaplan and Vakili (2014) adapted the latent Dirichlet allocation for the purpose of measuring degrees of novelty in collections of patents. Thorleuchter, van den Poel and Prinzie (2010) used simpler, similarity-based measures to investigate how one can extract new and useful ideas from unstructured text from research proposals. Netzer et al. (2012) used brand and word co-occurrence matrices extracted from an online automobile forum as input data for network analysis and multidimensional scaling, obtaining perceptual maps that describe the market position of the different brands. These contributions are interesting because, in their own respective ways, they seek to extract innovation-related information from collections of unstructured text. Two of the studies are particularly relevant for the present research: Netzer et al. (2012), because they use online communities as a data source, and Kaplan and Vakili (2014), because their central concept is the novelty of an idea.

Measuring the Presence and Quality of Ideas

In most studies, an idea refers to the initial outcome of a creative process that can be further developed into a proposal, prototype or tangible product (Wallas, 1926; Lubart, 2001; Dean et al., 2006; O'Quin & Besemer, 2006). Much research on creativity and fuzzy front-end innovation has focused on measuring the *quality* of ideas (Dean et al., 2006). In a typical study, a group of creatives generates ideas in a predefined domain and a group of assessors rates their quality on appropriate rating scales. Besemer (1998) and Besemer and O'Quin (1999) used this methodology to investigate which dimensions made the design of a chair particularly creative. Reinig, Briggs and Nunamaker (2007) compared different ways of scoring such data to evaluate the effectiveness of idea generation techniques. They recommended the number of good ideas generated as one of the best indicators of creativity and innovation. Kudrowitz and Wallace (2013) proposed a minimal set of rating scales (novelty, usefulness, feasibility) that have sufficient validity for an initial screening of the results of idea generation exercises. Kristensson, Gustafsson and Archer (2004) investigated whether ideas generated by expert users and ordinary users could compete with ideas generated by professional developers. Poetz and Schreier (2012) compared ideas generated by users in a crowdsourcing community and ideas generated by company professionals.

Most of the above studies were experimental; their instructions made sure that the outcomes of the idea generation exercises were in fact ideas. In such a situation, differences in the number and quality of the generated ideas are indeed the logical focus of the analysis. In an online community setting, however, most messages and comments will not contain any ideas at all. The few that do contain one or more ideas have to be identified *before* their quality can be assessed. This is generally only possible if people have ways of expressing ideas that manifest themselves in characteristic syntactic and lexical patterns, which are recognizable by human

© 2016 John Wiley & Sons Ltd

judges. And if these patterns are sufficiently stable, it should in principle be possible to train a computer algorithm to automatically detect ideas in collections of texts extracted from online communities.

In order to illustrate the task, consider two texts from the online community we used in our analysis (Table 1). The text on the left we interpret as containing an idea: here, a community member expresses a desirable outcome and a technical solution by which the outcome could be achieved. The text on the right we interpret as a chat between two community members, not containing an idea. We believe there is a clear difference in their innovation potential, and we also believe this difference is clearly recognizable from the different syntactic and lexical patterns in the texts. Hence, we argue that there is scope for a method that can automatically distinguish the underlying classes, separating idea texts from non-idea texts.

A Supervised-Learning Approach to Idea Detection

Machine learning is about teaching computers to recognize patterns. Typically, machine learning techniques are divided into two branches: supervised learning techniques (such as regression, discriminant analysis, decision trees, neural networks, and support vector machines) and unsupervised learning techniques (such as principal component analysis, cluster analysis and latent Dirichlet allocation). Unsupervised techniques are based on unlabelled data, i.e. categories of interest are not imposed on the studied data (Bao & Datta, 2014). This class of technique aims at *discovering* patterns in data and represents them in a low-dimensional form, often accompanied by visualizations that make them easier to interpret. Nevertheless, interpretation remains the job of the researcher. More importantly, it is not in the nature of unsupervised techniques to make *distinct* binary predictions (i.e. idea vs. non-idea) and they are therefore inherently *descriptive*. This is important, as all existing applications of text mining that we reviewed above (Thorleuchter, van den Poel & Prinzie, 2010; Netzer et al., 2012; Kaplan & Vakili, 2014; Antons et al., 2016) used unsupervised techniques. Hence, their methodology would not be applicable in a situation where the objective is to *distinctly* classify texts into one of two classes (idea text versus non-idea text).

Supervised learning techniques, on the other hand, are based on labelled data, i.e. a predefined set of categories (Bao & Datta, 2014). Here the value of a specific target variable (synonymous with a dependent or response variable) is predicted from the values of the input variables (synonymous with independent or predictor variables), given a model of the relationship between the input and target variables. The model can be statistical (e.g. a regression model) or algorithmic (e.g. a support vector machine), it can be linear or non-linear, and the target can be a continuous variable or a classification variable. The drawback of supervised learning techniques is that the model, whatever its nature, can only be estimated if a training sample exists in which the values of the target variable are known. Furthermore, an independent test sample is required to evaluate its performance in an unbiased manner.

Methods and Data

Training Data

The training data for our supervised idea detection task was extracted from the Lego online community LUGNET (the Lego User Group Network). The community was

Table 1. Idea Text and Non-Idea Text from Data Set

Idea Text	Non-Idea Text
'What I think would be really cool is a synchro-drive platform that can be controlled by one motor and therefore be watched by one rotation sensor. For example, motor forward drives the wheels to move the platform while motor reverse rotates the wheels. If this can be done then you could control and track your robot's position with a single output and a single input. That's a big IF though.:) Later, ##NAME## wrote: −− Did you check the website first?: ##COMPANY EMAIL ADDRESS##'	'If you hum a few bars, maybe. Seriously, I can't even whistle 300 baud. Although I had a roomie that could whistle 120. Remember 120 baud??? TI "portable" TTY#s with thermal paper printers?? -- ##COMPANY NAME## - ##EMAIL ADDRESS##. Mercator, the e-business transformation company fund Lugnet(tm): http://www.ebates.com/ ref: lar, 1/2 $$ to lugnet. Note: this is a family forum!'

© 2016 John Wiley & Sons Ltd

established in 1998 by a group of self-proclaimed Adult Fans of Lego (AFOLs). It offers AFOLs an online platform for sharing suggestions by hosting a web of individualized, member-created homepages, accessing a variety of topical and geographical Lego User Groups (known as LUGs), sharing information about Lego products and Lego-related resources on the Internet, and finally, selling, buying and trading Lego sets and elements by providing a more efficient 'integrated' marketplace (Antorini, 2007). AFOLs are known for their ability to develop innovations (Nørskov, Antorini & Jensen, 2016), and they have generated new products and new product lines and created new market opportunities for Lego (Antorini, Muñiz & Askildsen, 2012). The AFOLs' innovations have created value both for the user innovators and the company. Therefore this particular Lego community is relevant for our study of idea generation.

To generate the target variable, we extracted a random selection of 3,000 messages from the LUGNET news server. Two individuals were recruited as idea raters and instructed to read each text and evaluate whether it contained suggestions about products, improvements, or business opportunities. If it did, the raters were instructed to assign a target value of $y = 1$ to the text. If it did not, the raters were instructed to assign a target value of $y = 0$ to the text. After the rater training was completed, both raters independently classified the 3,000 texts. Rater 1 classified 8.73% of the texts as containing at least one idea (corresponding to 264 idea texts and 2,736 non-idea texts). Rater 2 classified 6.90% as containing at least one idea (207 idea texts and 2,793 non-idea texts). The raters agreed on 137 idea texts and 2,666 non-idea texts. (The remaining 197 texts were later omitted from the analysis.)

Cohen's kappa was calculated as a measure of inter-rater reliability. Kappa is often interpreted using the following thresholds: $\kappa < 0$, poor; $0 < \kappa \leq 0.20$, slight; $0.20 < \kappa \leq 0.40$, fair; $0.4 < \kappa \leq 0.60$, moderate; $0.60 < \kappa \leq 0.80$, substantial; and $0.80 < \kappa \leq 1$, almost perfect (Cohen, 1960; Landis & Koch, 1977). In the present case, the result was $\kappa = 0.55$ (± 0.08 at $\alpha = 0.05$), a value that would normally be regarded as moderate. However, the theoretical maximum of kappa depends on the marginal distributions of the codes assigned by the raters (von Eye & von Eye, 2008). In the present case, the marginal distributions differed so that the theoretical maximum of kappa was only $\kappa_{\max} = 0.87$. Hence, the observed value of kappa was 63% of its maximum value, moving it into a range that can be regarded as substantial.

Data Pre-Processing

Before unstructured texts can be used in machine learning, they have to be pre-processed. We removed all punctuation marks, numbers and additional white spaces from the 3,000 LUGNET posts we had extracted, converted all uppercase letters to lowercase letters, and removed citations of previous posts to which the texts responded. We experimented with stopword removal, creating versions of the data set in which stopwords were and were not removed. In addition, we identified all possible n-grams up to an order of $n = 3$ in the texts. N-grams are sequences of words that carry additional meaning (e.g., 'this' is a unigram, 'I like' is a bigram and 'this is nice' is a trigram). The effect of using n-grams is that words are allowed to interact, creating additional nuances of meaning (Zanasi, 2007). In the present analysis, we experimented with different orders of n-grams, creating versions of the data set in which only unigrams were included as terms, where unigrams and bigrams were included, and where unigrams, bigrams and trigrams were included.

All unique terms were counted and transformed into a 'bag-of-words' representation where texts were represented as rows (observations) and terms as columns (variables). We experimented with the term weighting using the two most common schemes: term occurrences (where weights equal the raw counts of terms in a given document) and binary term occurrences (where a weight of 1 indicates that the term occurs at least once in a given document and 0 indicates that it does not). Finally, we reduced the bag-of-words representation to a computationally more feasible size by setting a sparsity threshold, eliminating all terms that occurred in a lower proportion of the texts than the defined threshold. Exactly how many terms to exclude is debateable and requires careful consideration. Tirunillai and Tellis (2014), for example, used a sparsity threshold of 2% and Antons et al., (2016) used a sparsity threshold of 0.1%. In the present analysis, we experimented with sparsity thresholds of 2.5%, 1% and 0.25%.

Partitioning into Training and Validation Sets

The data were partitioned into a training set (75% of the texts) and an independent validation set (25% of the texts). We used stratified random sampling with stratification on the target variable, resulting in comparable target variable distributions in both sets. In the training set, we estimated altogether 252

© 2016 John Wiley & Sons Ltd

alternative classification models that differed in terms of the underlying data representation (stopword removal, sparsity threshold, order of *n*-grams included, term weighting scheme; see above) and the tuning parameters of the classification technique (see below). In the validation set, we compared the performance of the 252 classifiers on previously unseen data and selected the model with the highest performance.

Classification Technique

The choice of a particular family of classification techniques (e.g. neural networks, nearest neighbour classifiers, decision trees, naïve Bayes classifiers, support vector machines; see Witten & Frank, 2005; Han & Kamber, 2006; Hastie, Tibshirani & Friedman, 2008; Linoff & Berry, 2011) will impact the types of patterns a classifier will be able to recognize. However, this strongly depends on the nature of the input data. In text mining applications, the input data are usually high-dimensional, consisting of as many variables as there are unique terms (typically several thousand). Classification techniques that perform well under conditions of high dimensionality are therefore a necessity in text mining. Table 2 summarizes the results of studies that compared the performance of different families of classification techniques, with an emphasis on text mining applications.

In the majority of these performance comparisons, support vector machines (Cortes & Vapnik, 1995) were the most powerful technique. Similar to discriminant analysis, a linear support vector machine technique tries to find a hyperplane that separates the target classes in the space of the input variables. However, the optimization criterion is the *width* of the margin by which the classes are separated: unlike in discriminant analysis, the estimation of the parameters of the hyperplane depends exclusively on the observations (= the support vectors) that lie on the margins around the hyperplane. If the target classes are not linearly separable, a soft-margin constant *C* can be introduced that determines how many observations are allowed to lie within the margins and how far they are allowed to do so. *C* can be understood as a penalty term: the higher its value, they stronger the penalty on margin violations and the lower the flexibility of the model (for details, see Cortes & Vapnik, 1995; Ben-Hur & Weston, 2010). *C* is a hyper-parameter with a data-dependent optimum; it cannot usually be generalized from one modelling context to another. A suitable value has to be found computationally, for example through a grid search over a range of possible values. In the present analysis, we experimented with values of 1e-05, 1e-04, 0.001, 0.01, 0.1, 1 and 10.

Like most classification techniques, support vector machines can be trained most effectively when the distribution of the target variable is approximately equal in the training set (Menardi & Torelli, 2014). In the present case, however, the distribution was unbalanced: 4.9% idea texts compared to 95.1% non-idea texts. Although this is not extreme (ratios as high as 1:100, 1:1,000, or 1:10,000 are not uncommon in real-world classification problems; see Weiss & Provost, 2001), we tried to improve the learning conditions by using a particular bootstrap aggregation approach known as 'under-bagging' (see Breiman, 1996). In each bootstrap replication, all idea texts are used but combined with a different subsample of non-idea texts with the same sample size as the number of idea texts. The results were then aggregated using different voting schemes (Galar et al., 2012); for example, majority voting, where each text is assigned to the class of the target variable that the majority of the bootstrap classifiers predict, or unanimous voting, where a text is only assigned to the idea class if all bootstrap classifiers in the ensemble agree.

Assessment of Classification Performance

The performance of the competing classification models were compared using measures derived from signal detection theory (see Stanislaw & Todorov, 1999; Witten & Frank, 2005). Signal detection theory is particularly applicable to binary classification tasks where the presence of a particular target event is of interest (the signal; in this case presence of one or more ideas in a text) and its absence can be regarded as noise. All classification performance measures are derived from the confusion matrix, a cross-tabulation of the classification results against the true class membership of the texts. In our case, true positives (TP) are idea texts that were correctly identified as idea texts by the classifier. True negatives (TN) are non-idea texts correctly identified as non-idea texts. False positives (FP) are non-idea texts that were incorrectly classified as idea texts. False negatives (FN) are idea texts that were incorrectly classified as non-idea texts. Based on these counts, numerous measures can be calculated that quantify different aspects of classification performance. We will use five of these: recall, precision, the F_1-measure, classification accuracy and Cohen's kappa. Recall (also known as sensitivity, true positive rate, or hit rate) is the proportion of ideas that the classifier correctly detected. Precision (also known as positive predictive value or one minus the false discovery rate) is the proportion

© 2016 John Wiley & Sons Ltd

Table 2. Identified Studies that Were Aimed at Comparing Supervised Machine Learning Techniques for High-Dimensional Datasets

Source	Study Type	Data Type	Ranked Classification Performance
Amancio et al. (2014)	Comparative study	Artificial data	(1) Support vector machines (2) Nearest neighbour (3) Decision trees (4) Neural networks
Bijalwan et al. (2014)	Comparative study	Text data	(1) Nearest neighbours (2) Naïve Bayes (3) Alternative technique
Khan et al. (2010)	Review of supervised learning techniques for text mining		(1) Support vector machines (2) Nearest neighbour (3) Naïve Bayes (4) Neural networks (5) Decision trees
Ye, Zhang & Law (2009)	Comparative study	Text data	(1) Support vector machines (2) Alternative technique (3) Naïve Bayes
Lai (2007)	Comparative study	Text data	(1) Support vector machines (2) Naïve Bayes (3) Nearest neighbour
Kotsiantis (2007)	Review of supervised learning techniques in general		(1) Support vector machines (2) Neural networks (3) Decision trees (4) Nearest neighbour (5) Naïve Bayes (6) Alternative technique
Zhang, Zhu & Yao (2004)	Comparative Study	Text data	(1) Support vector machines (2) Decision trees (3) Alternative technique (4) Naïve Bayes (5) Nearest neighbour
Pang, Lee & Vaithyanathan (2002)	Comparative Study	Text data	(1) Support vector machines (2) Alternative technique (3) Naïve Bayes
Drucker, Wu & Vapnik (1999)	Comparative Study	Text data	(1) Decision trees (2) Support vector machines (3) Alternative technique
Joachims (1998)	Comparative Study	Text data	(1) Support vector machines (2) Nearest neighbour (3) Alternative technique (4) Decision trees (5) Naïve Bayes

of texts classified as ideas that are in fact ideas. The F_1-measure is a compromise between recall and precision, based on their harmonic mean. As a model evaluation criterion, it is particularly useful in information retrieval tasks as it represents a 'fair' trade-off between the objectives of maximizing the true positives and minimizing the false positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

© 2016 John Wiley & Sons Ltd

$$F_1 = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

Recall, precision and the F_1-measure disregard all true negatives and are unaffected by the ability of a classifier to screen out true negatives. We will therefore report two additional measures that are more symmetric in this regard. Classification accuracy is the total proportion of correctly classified texts. Cohen's kappa (which we already used as a measure of inter-rater reliability; see above) is a corrected version of classification accuracy that takes the probability of chance agreement into account:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\kappa = \frac{\text{Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}} \quad (5)$$

External Validity: Classification Performance in an Independent Test Set

The decisive test of a the predictive power of a classifier is its performance in an independent test set that consists of completely new data (Hastie, Tibshirani & Friedman, 2008). The data in this test set should be from the same real-world domain for which the classifier was trained but should not in any way have been available during the training and tuning process or the selection of the final model. To construct an independent test set, we used our final classifier to extract a balanced sample of 500 new texts from the LUGNET newsgroups: 250 texts which the final classifier labelled as ideas and 250 texts it labelled as non-ideas.

The texts in the new test set were then independently classified by five different raters recruited from the crowdsourcing service *Crowdflower*. Based on their responses, we constructed a new target variable that took the value $y = 1$ if at least three out of five raters classified the text as containing at least one idea, and $y = 0$ otherwise. Classification performance in the test set was again measured in terms of recall, precision, the F_1-measure, classification accuracy and Cohen's kappa.

Results

Training and Tuning Procedure.

Throughout the analysis, we used linear support vector machines as the basic learning algorithm. Altogether 252 classifiers with different input data settings and hyper-parameters were trained on the training set. The best combination of input data settings and hyper-parameters was determined based on the F_1-measure obtained in the validation set. The tuning proceeded in two steps:

1. The best combination of soft-margin constant C, sparsity threshold, term weighting scheme, n-gram generation and stopword removal setting was determined based on the mean F_1-measure obtained in the validation set. C values were set to 1e-05, 1e-04, 0.001, 0.01, 0.1, 1 and 10, respectively. Ten replications were used in the under-bagging loops. A text was classified as an idea text if and only if all ten bootstrap classifiers in the ensemble agreed that the text belonged to the idea class.
2. The best combination of settings and tuning parameters from Step (1) was frozen. The optimal number of replications in the under-bagging loop was then determined based on the F_1-measure obtained in the validation set. The maximum number of classifiers in the ensemble was set to 25. Again, a text was classified as an idea text if and only if all bootstrap classifiers in the ensemble agreed that the text belonged to the idea class.

Classification Performance in the Validation Set

Classification performance in the validation set is summarized in Fig. 1 (grouped by input data and hyper-parameter settings). The soft-margin constant leading to the highest overall classification performance was $C = 0.01$ (Table 3). For this specific value, the best n-gram configuration was $n = 3$, with unigrams, bigrams and trigrams included as terms in the bag-of-words representation. The best stopword removal setting was not to remove stopwords. The best sparsity threshold was 0.25%, resulting in a vocabulary of 9,152 unique terms. The best term weighting scheme was term occurrences. The best number of bootstrap classifiers in the ensemble was 14.

The performance of the final model in the validation sample was very satisfactory (Table 4). Recall was 0.79 and precision was 0.42, combining a high ability to detect true ideas with a medium rate of false discoveries. The resulting F_1-measure was 0.55, a value that would be regarded as 'good' by most text miners. Overall classification accuracy was 0.94 and Cohen's kappa was 0.51 (77% of its theoretical maximum, given the marginal distributions), indicating that the classifier was also moderately effective in screening out non-ideas.

© 2016 John Wiley & Sons Ltd

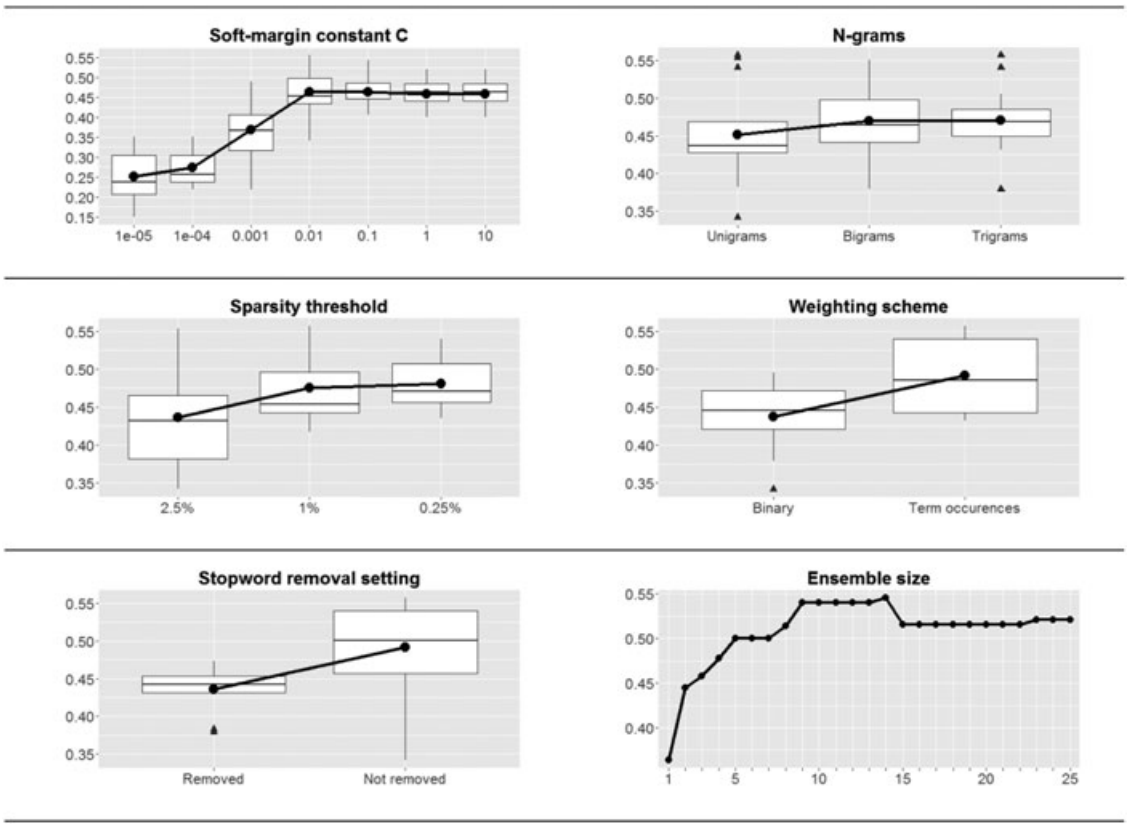


Figure 1. Distribution of Classification Performance (F_1-Measure) in the Validation Set as a Function of Soft-Margin Constant C, Order of N-grams Included in the Representation, Sparsity Threshold, Weighting Scheme, Stopword Removal Setting, and Ensemble Size

Table 3. Classifier Tuning with Respect to Cost (C)

C	TP	TN	FP	FN	F_1	Accuracy	Recall	Precision	Kappa
0.00001	5	638	28	29	0.15	0.92	0.15	0.15	0.11
0.0001	14	605	61	20	0.26	0.88	0.41	0.19	0.20
0.001	22	629	37	12	0.47	0.93	0.65	0.37	0.44
0.01	***27***	***627***	***39***	***7***	***0.54***	***0.93***	***0.79***	***0.41***	***0.51***
0.1	27	616	50	7	0.49	0.92	0.79	0.35	0.51
1	27	616	50	7	0.49	0.92	0.79	0.35	0.51
10	27	616	50	7	0.49	0.92	0.79	0.35	0.51

Abbreviations: TP = True ideas, TN = True non-ideas, FP = False ideas, FN = False non-ideas

Classification Performance in the Independent Test Set

The performance of the final model in the independent test set was excellent (Table 4). Recall was 0.72 and precision was 0.91, combining a high ability to detect true ideas with a low rate of false discoveries. The resulting F_1-measure was 0.81, an even better value than the performance achieved in the validation sample. Overall classification accuracy was 0.78 and Cohen's kappa was 0.56 (76% of its theoretical maximum), indicating that the classifier was as effective in screening out the true non-ideas as it had been in the validation sample.

© 2016 John Wiley & Sons Ltd

Table 4. Classifier Performance on Validation Set and External (Crowd) Test Set

Set	TP	TN	FP	FN	F_1	Accuracy	Recall	Precision	Kappa
Validation	27	628	38	7	0.55	0.94	0.79	0.42	0.51
Test	228	162	22	88	0.81	0.78	0.72	0.91	0.56

Abbreviations: TP = True ideas, TN = True non-ideas, FP = False ideas, FN = False non-ideas

Discussion and Implications

We offer a novel method for detecting ideas in online communities via machine learning and text mining. Our study contributes to the innovation management literature by extending the current knowledge on the automation of idea identification by applying supervised learning techniques. It also brings interesting insights to researchers and a new operational tool for managers working at the fuzzy front-end of innovation. We argue that the central premise for developing such a method is that an idea is manifested on a syntactical level, following a specific pattern, when transformed from thought to written text. This pattern should be recognizable by two human raters rating independently of one another. Our results show that the kappa agreement between the two raters recruited for this study was *substantial* according to the applied benchmark scale. Therefore the proposed method for collecting texts and identifying ideas is considered useful for building a reliable target variable.

A reliable target variable can be used for training a machine learning classifier, and the results of our training and testing procedure are reported in Table 4. The next natural question is: Are these results *satisfactory*? To answer this question, we assessed the obtained kappa measures in relation to the benchmark scale we already applied and found that the kappa measures were *substantial* for the validation set as well as for the external set. Further, when comparing our results with results from similar idea detection studies, Thorleuchter, van den Poel and Prinzie (2010) developed a measure that could be used to find new ideas amongst a database of patents. When a human rater was asked to evaluate the identified ideas, precision was 0.40 and recall was 0.25. In a similar study where the idea database was the entire World Wide Web, Thorleuchter and van den Poel (2013) obtained an F_1-measure ranging from 0.29 to 0.38. We obtained an F_1-measure ranging from 0.55 to 0.81. Finally, most of the results reported in the studies in Table 2 obtained accuracy measures in the range 0.85 to 0.95. This is similar to the accuracy we obtained on the validation set. However, our accuracy on the external set was notably lower. We speculate that this is because most of the identified studies did not measure performance on a third external set, but either applied a two-split strategy or a cross-validation strategy, which can yield optimistic results (Hastie, Tibshirani & Friedman, 2008). Despite the low accuracy on the external set, we consider our results satisfactory.

In the above paragraph, we assessed the results in relation to previous studies by assessing kappa, the F_1-measure and accuracy. However, judging *whether* the results we obtained are satisfactory is a decision that should not be made in relation to theoretical benchmark scales and previous studies alone, but should also relate to the practical implications of the method. Therefore, if we turn our attention to the recall measure obtained on the external set (0.72), the practical implications of the proposed method is that it can be expected to identify 720 out of 1,000 idea texts. The results also show a precision value of 0.91. This implies that when a trained classifier extracts 1,000 ideas from a similar online community, 910 of the texts will be true ideas and 90 of the texts will be non-ideas. These results are interesting because: (1) it is possible that artificial intelligence in terms of machine learning algorithms can learn and recognize abstract entities as ideas; and (2) the method can be used as a pre-filter, which can be used for extracting texts *before* assigning human raters to coding. Such a method would be useful for studying the quality of ideas generated in online communities, with the pre-filter applied to data from an online community of the researcher's own choice. This means that if the researcher wants to study 100 ideas, the researcher would have to extract approximately 110 texts identified as ideas by the method, and recruit two human raters to verify which of the texts are in fact ideas.

Innovation practitioners may benefit from our method, as well. The proposed method could potentially allow firms to reduce the cost of idea identification in online communities. The two raters recruited for this study were paid US$6,500. They assessed 3,000 texts each and they identified 137 ideas. This corresponds

© 2016 John Wiley & Sons Ltd

to a price of US$47.45 per idea and US$2.17 per text assessment. The costs of identifying 100 ideas would then sum to US$4,745 if no pre-filtering were applied. If, on the other hand, our method was applied as a pre-filter, the identification of 100 ideas would cost US$238 in total. This corresponds to only 5% of the cost without the pre-filtering method. The firm would, however, need to accept the 28% loss of true ideas (Recall =0.72).

The loss of 28% of the ideas is a consequence of how our classifier is tuned. By tuning with respect to the F_1-measure, the implicit assumption was made that the optimal solution is the one that balances precision and recall. This trade-off is what the F_1-measure seeks to balance, and by making the choice to tune with respect to the F_1-measure, all choices that were made throughout the tuning process were in favour of the F_1-measure. However, one can imagine cases where a researcher or a firm would favour a solution that found as many ideas as possible at the cost of lower precision. If, for example, ideas are rarer than in our case, or the costs of doing manual classifications are low, it might be preferable to choose a solution that favours recall rather than the F_1-measure. In relation to this discussion, it might be relevant to mention that in all our testing the maximum recall obtained was 0.94. As a consequence of this high recall, precision would drop to 0.20. We report this because favouring recall might be better from a practitioner's viewpoint, as it would incur a loss of only 6% of the ideas and require reading five texts to find one idea.

In our case, the two raters classified 4.57% (137) of the texts as ideas; they disagreed on 6.57% (197) of the texts; and they classified 88.86% (2,666) of the texts as non-ideas. These numbers are interesting because they tell us that ideas may be a rare kind of information in an online community. However, the rarity of ideas does not mean that they are more interesting or relevant than other types of information. The relevance of a particular type of information can only be assessed by those persons or organizations that absorb the information.

For example, the two texts displayed in Table 5 contain an idea text that proposes a new product and a non-idea text that is interpreted as spam or advertisement. For the researcher who wishes to investigate the potential of online communities for new product development, the idea text would be interesting. For the researcher who wishes to investigate spam infiltration in online communities, the non-idea may be interesting. In this paper, we have developed a method that makes it possible to detect ideas, but one should not neglect the fact that online communities contain other types of information that might be interesting for specific purposes. Therefore, future research could specify other types of information (e.g. product-related problems, purchasing experiences, etc.) that might be of interest, and develop methods that can identify such information. Together with our method for detecting ideas, such a set of methods could pave the way for a new stream of research by innovation and marketing management scholars that could help firms learn how to better engage with, collaborate with, and/or integrate online communities into firms' new product development and innovation activities.

Limitations and Suggestions for Future Research

The main limitation of this study is that our method is only tested on texts related to one specific community (LUGNET) and one specific product domain (toys). Future research should therefore focus on validating the method on texts from other online communities within the same and other product domains. This would require the creation of similar training sets,

Table 5. Idea Text and Non-Idea Text Identified by Two Human Raters

Idea Text	Non-Idea Text
'Hmm. I wonder if ##COMPANY NAME## thought about making a combo-pack of something like #### Shark Cage Cove and #### Cross Bone Clipper. If they had included an additional instruction set, to build a shipwreck search-and-salvage (or the academic equivalent of salvage), that would have been great! ##NAME##'	'I'm sure a lot of you are the same way. This is why I'm telling you about my part trade site at: http://members.xoom.com/WDS/trade/index.html This is not an auction and it is not for profit (trades only, no sales). I do this only to find unneeded / unwanted ##COMPANY NAME## new homes. Friend to every yellow ##COMPANY NAME##... ##NAME##'

© 2016 John Wiley & Sons Ltd

which is probably the biggest obstacle given that manual coding can be costly. For this study, two Master's students of innovation management were recruited to rate the texts. As mentioned, the two raters were paid approximately US$6,500 altogether for evaluating the texts. We consider these costs high, and we suggest that future research focuses on developing methods that can lower the costs of doing such manual evaluations that require human intelligence. By our validation study we have shown that crowdsourcing is a suitable solution to this problem.

Another limitation is that the 197 texts on which our raters disagreed were omitted from the machine learning procedure. This problem could potentially be avoided by designing the evaluating task differently. Instead of asking the raters to evaluate the texts in a binary fashion, one could have asked for a continuous response, e.g. '*does the text contain one or several ideas?*' The response scale could then have been 1 for 'absolutely not' and 10 for 'absolutely'. This would in return have required the classification task to be framed as a regression task.

Yet another limitation is the performance scores of the classifier on the training validation set. Despite the high overall accuracy and substantial kappa measures, the performance of the classifier on precision, recall and F_1-measure was mediocre with respect to the idea class. We hypothesize that this limitation is primarily a consequence of having too few ideas vs. to many non-ideas available for training. We highlight the importance of balancing the target distribution in future studies *or* gathering more training data.

A final limitation is that we tested only one machine learning algorithm, namely the linear support vector machine. This decision was made based on a literature review and preliminary testing of the algorithms of naïve Bayes, decision trees, nearest neighbours, neural networks, and radial-basis support vector machines. We chose to omit these algorithms because we found it essential to introduce and explain the rationale behind the method so that it would be understandable by people with little knowledge of machine learning and text mining. We could have chosen to also show the performance of these classifiers, but it was our concern that introducing additional algorithms would shift the focus to the technical algorithms rather than the rationale behind utilizing machine learning for research. Future research should thus focus on confirming or rejecting the choice of the linear support vector machine as the best algorithm, and, in addition, test algorithms that also allow for explaining the phenomenon rather than simply predicting it.

If our method can be further developed and its scope extended to include not only ideas related to toys, but ideas related to any product domain, innovation and creativity researchers can start asking and answering a new range of research questions. These could be related to the social systems in online forums, social media or the blogosphere and their relation to the offline world. Is there a relationship between online idea generation, the industry or firm, and the performance of that same industry or firm? Is the innovation performed in a given industry reflected in online conversations? Or is the causality reversed so that online ideation serves as a catalyst for industry and firm innovation?

Conclusion

We propose a method for automatically identifying ideas written as text in online communities. Our results support the claim that artificial intelligence has reached a state where it can add a new dimension to key tasks of innovation activities. The method was developed based on supervised machine learning and text mining techniques. The machine learning task was defined as a binary classification task and 3,000 texts were extracted from an online community where the topic of interest is toys. We used a linear support vector machine to test whether a machine learning classifier of this nature could learn the pattern of ideas written as text. The comparison between performance on the validation set and performance on the external test set showed minor signs of over-fitting, which supports the reliability of the classifier and the method. We hope that our contribution inspires other researchers to develop methods of a similar nature, and so aid to develop this particular method of automatic idea detection.

Acknowledgments

The authors would like to thank The Foundation for Research and Levy on Agricultural Products in Norway for funding the time that was spend on developing this paper.

References

Amancio, D.R., Comin, C.H., Casanova, D., Travieso, G., Bruno, O.M., Rodrigues, F.A. and da Fontoura Costa, L. (2014) A Systematic Comparison of Supervised Classifiers. *PloS One*, 9, 1–4.

Antons, D., Kleer, R. and Salge, T.O. (2016) Mapping the Topic Landscape of *JPIM*, 1984–2013: In Search of Hidden Structures and Development

© 2016 John Wiley & Sons Ltd

Trajectories. *Journal of Product Innovation Management*, 33, 726–49.

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

Antorini, Y.M., Muñiz, A.M. and Askildsen, T. (2012) Collaborating with Customer Communities: Lessons from the Lego Group. *MIT Sloan Management Review*, 53, 73–95.

Bao, Y. and Datta, A. (2014) Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science*, 60, 1371–91.

Ben-Hur, A. and Weston, J. (2010) A User's Guide to Support Vector Machines. *Methods in Molecular Biology*, 609, 223–39.

Besemer, S.P. (1998) Creative Product Matrix Analysis: Testing the Model and a Comparison among Products – Three Novel Chairs. *Creativity Research Journal*, 11, 333–46.

Besemer, S.P. and O'Quin, K. (1999) Confirming the Three-Factor Creative Product Analysis Matrix Model in an American Sample. *Creativity Research Journal*, 12, 287–96.

Bijalwan, V., Kumar, V., Kumari, P. and Pascual, J. (2014) KNN Based Machine Learning Approach for Text and Document Mining. *International Journal of Database Theory and Application*, 7, 61–70.

Björk, J. and Magnusson, M. (2009) Where Do Good Innovation Ideas Come From? Exploring the Influence of Network Connectivity on Innovation Idea Quality. *Journal of Product Innovation Management*, 26, 662–70.

Breiman, L. (1996) Bagging Predictors. *Machine Learning*, 24, 123–40.

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

Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20, 273–97.

Dahlander, L. and Piezunka, H. (2015) Distant Search, Narrow Attention: How Crowding Alters Organizations Filtering of Suggestions in Crowdsourcing. *Academy of Management Journal*, 58, 856–80.

Dean, D.L., Hender, J.M., Rodgers, T.L. and Santanen, E.L. (2006) Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7, 646–98.

di Gangi, P.M., Wasko, M.M. and 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, 213–28.

Drucker, H., Wu, D. and Vapnik, V.N. (1999) Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks*, 10, 1048–54.

Foss, N.J., Frederiksen, L. and Rullani, F. (2015) Problem-Formulation and Problem-Solving in Self-Organized Communities: How Modes of Communication Shape Project Behaviors in the Free Open-Source Software Community. *Strategic Management Journal*, doi:10.1002/smj.2439.

Füller, J., Bartl, M., Ernst, H. and Mühlbacher, H. (2006) Community Based Innovation: How to Integrate Members of Virtual Communities into New Product Development. *Electronic Commerce Research*, 6, 57–73.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 42, 463–84.

Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, CA.

Hastie, T., Tibshirani, R. and Friedman, J. (2008) *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, 2nd edn. Springer, Stanford, CA.

Henkel, J. (2006) Selective Revealing in Open Innovation Processes: The Case of Embedded Linux. *Research Policy*, 35, 953–69.

Hertel, G., Niedner, S. and Herrmann, S. (2003) Motivation of Software Developers in Open Source Projects: An Internet-Based Survey of Contributors to the Linux Kernel. *Research Policy*, 32, 1159–77.

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

Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Nédellec, C. and Rouveirol, C. (eds.), *Proceedings of 10th European Conference on Machine Learning (ECML-98)*. Chemnitz, Germany, pp. 137–42.

Kaplan, S. and Vakili, K. (2014) The Double-Edged Sword of Recombination in Breakthrough Innovation. *Strategic Management Journal*, 36, 1435–57.

Khan, A., Baharudin, B., Lee, L.H. and Khan, K. (2010) A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1, 4–20.

Kotsiantis, S.B. (2007) Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31, 249–68.

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

Kudrowitz, B.M. and Wallace, D. (2013) Assessing the Quality of Ideas from Prolific, Early-Stage Product Ideation. *Journal of Engineering Design*, 24, 120–39.

Lai, C.-C. (2007) An Empirical Study of Three Machine Learning Methods for Spam Filtering. *Knowledge-Based Systems*, 20, 249–54.

© 2016 John Wiley & Sons Ltd

Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33, 159–74.

Linoff, G. and Berry, M. (2011) *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 3rd edn. Wiley, Indianapolis, IN.

Lubart, T.I. (2001) Models of the Creative Process: Past, Present and Future. *Creativity Research Journal*, 13, 295–308.

Menardi, G. and Torelli, N. (2014) Training and Assessing Classification Rules with Imbalanced Data. *Data Mining and Knowledge Discovery*, 28, 92–122.

Netzer, O., Feldman, R., Goldenberg, J. and Fresko, M. (2012) Mine Your Own Business: Market-Structure Surveillance through Text Mining. *Marketing Science*, 31, 521–43.

Nørskov, S., Antorini, Y.M. and Jensen, M.B. (2016) Innovative Brand Community Members and their Willingness to Share Ideas with Companies. *International Journal of Innovation Management*, 20 .1650046

Ogawa, S. and Piller, F.T. (2006) Reducing the Risks of New Product Development. *MIT Sloan Management Review*, 47, 65–71.

O'Quin, K. and Besemer, S.P. (2006) Using the Creative Product Semantic Scale as a Metric for Results-Oriented Business. *Creativity and Innovation Management*, 15, 34–44.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, pp. 79–86.

Poetz, M.K. and Schreier, M. (2012) The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas? *Journal of Product Innovation Management*, 29, 245–56.

Reinig, B., Briggs, R. and Nunamaker, J. (2007) On the Measurement of Ideation Quality. *Journal of Management Information Systems*, 23, 143–61.

Sawhney, M. and Prandelli, E. (2000) Communities of Creation: Managing Distributed Innovation in Turbulent Markets. *California Management Review*, 42, 24–54.

Stanislaw, H. and Todorov, N. (1999) Calculation of Signal Detection Theory Measures. *Behavior Research Methods, Instruments & Computers*, 31, 137–49.

Thorleuchter, D. and van den Poel, D. (2013) Web Mining Based Extraction of Problem Solution Ideas. *Expert Systems with Applications*, 40, 3961–9.

Thorleuchter, D., van den Poel, D. and Prinzie, A. (2010) Mining Ideas from Textual Information. *Expert Systems with Applications*, 37, 7182–8.

Tirunillai, S. and Tellis, G.J. (2014) Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, 51, 463–79.

von Eye, A. and von Eye, M. (2008) On the Marginal Dependency of Cohen's κ. *European Psychologist*, 13, 305–15.

von Krogh, G., Spaeth, S. and Lakhani, K.R. (2003) Community, Joining, and Specialization in Open Source Software Innovation: A Case Study. *Research Policy*, 32, 1217–41.

Wallas, G. (1926) *The Art of Thought*. Solis Press, New York.

Wasko, M.M. and Faraj, S. (2005) Why Should I Share? *Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. MIS Quarterly*, 35–57.

Weiss, G.M. and Provost, F. (2001) *The Effect of Class Distribution on Classifier Learning: An Empirical Study*. (Technical report No. ML-TR-44)Department of Computer Science, Rutgers University, Newark, NY.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, CA.

Ye, Q., Zhang, Z. and Law, R. (2009) Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches. *Expert Systems with Applications*, 36, 6527–35.

Zanasi, A. (2007) *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. WIT Press, Southampton.

Zhang, L., Zhu, J. and Yao, T. (2004) An Evaluation of Statistical Spam Filtering Techniques. *ACM Transactions on Asian Language Information Processing*, 3, 243–69.

Kasper Christensen (kasper.christensen@nofima.no, kasper2304@gmail.com) is a PhD student at the Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences. He has a Masters degree in Marketing from the Business and Social Sciences Faculty at Aarhus University. His main research interests are innovation management and marketing management in combination with applied text mining and applied machine learning.

Sladjana Nørskov (norskov@mgmt.au.dk) is a postdoctoral researcher at the Department of Management, Aarhus University. She received her PhD from Aarhus School of Business. Her research interests include innovation management, organization development, user-centered innovation processes and new organizational forms. Her work appears in journals such as *Information Technology & People*, *Journal of Consumer Marketing* and *European Journal of Innovation Management*.

Lars Frederiksen (l.frederiksen@mgmt.au.dk) is Professor in the Department of Management, Aarhus University, Denmark, where he serves as Head of Research and

© 2016 John Wiley & Sons Ltd

Talent. He was awarded his PhD from Copenhagen Business School, Denmark, and then worked at Imperial College Business School, London. Lars specializes in the management of innovation and technology with particular emphasis on innovation strategies, knowledge creation and search, user innovation in communities, and innovation in project-based organizations. Lars' work appears in journals such as *Organization Science*, *Academy of Management Journal*, *Strategic Management Journal*, among others.

Joachim Scholderer (js@econ.au.dk) is a Professor in the Department of Economics and Business Economics at Aarhus University in Denmark. He also holds a position as Research Director at CCRS at the University of Zürich. Joachim's work has been published in *Journal of Business Research*, *Research Policy*, *Risk Analysis*, among others.

© 2016 John Wiley & Sons Ltd

Paper 2

Nina Veflen Olsen and Kasper Christensen

Social media, new digital technologies and their potential application in sensory and consumer research (2015). Current Opinion in Food Science. vol. 3, pp. 23-26

The paper is reprinted with permission from Elsevier

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Food Science

Social media, new digital technologies and their potential application in sensory and consumer research

Nina Veflen Olsen[1,2] and Kasper Christensen[1,2]

New digital technologies have changed the way people communicate and opened up for new ways of interacting with consumers via social media. This article reviews findings from recent investigations and present the opportunities and challenges social media offers for sensory and consumer science. After defining social media and giving a short overview of the different medium that exists, the focus will be on two aspects of specific interest: crowdsourcing and communication of health and food safety.

Addresses
[1] Aarhus University, Denmark
[2] Nofima, Norway

Corresponding author: Olsen, Nina Veflen (nvol@nofima.no, nvo@auhe.au.dk)

Current Opinion in Food Science 2015, **3**:23–26

This review comes from a themed issue on **Sensory science and consumer perception**

Edited by **Paula Varela-Tomasco**

http://dx.doi.org/10.1016/j.cofs.2014.11.006

2214-7993/© 2014 Elsevier Ltd. All rights reserved.

Introduction

New digital technology has made the exchange of user generated content on internet possible and turned the web into a very popular social medium. Facebook alone has over one billion active users and many people spend today more than one third of their waking day consuming social media content [1]. People share life stories and personal opinions in blogs, write short comments on Twitter, chat with their friends on Facebook, post pictures in Instagram and Flickr, watch other peoples' videos on You Tube and send small snaps of what they are doing on Snapchat. They share information and express their emotions. They tell life stories and give advice. They brag and they complain. People are no longer only passive consumers of professional internet content; they participate actively in creating and sharing their own content. This interactivity creates a lot of opportunities and challenges, so also for sensory and consumer science. Social media makes global, one-to-one communication easier and cheaper than ever, makes the voice of the consumer much stronger, and allows a dissatisfied costumer not only to complain to her friends but to post negative comments to millions of people [2].

The aim of this paper is to review recent literature and present the opportunities and challenges social media offers for sensory and consumer science. After defining the term social media and giving a short overview of the different types, the focus will be on two specific aspects: crowdsourcing and communication of health and food safety

Social media

Social media is defined as 'a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and allow the creation and exchange of User Generated Content [3]'. Technical functions as Adobe Flash (for animation and audio/video steam updates), Really Simple Syndication (for frequent updates of blogs and headlines) and Asynchronous Java Script (for update of web content without interfering with the interface of the whole page) made it possible for unprofessional users to make their own content publicly available over the internet (see [4] for a definition of User Generated Content).

Social media comes in many different forms: blogs, micro blogs (Twitter), collaborative projects (Wikipedia), social networking sites (Facebook), content communities (YouTube), virtual social worlds (Second Life), virtual game worlds (World of Warcraft). Different attempts have been made to classify social media according to theories like: Social Presence Theory, which states that medium differs in the degree that acoustic, visual and physical contact can be achieved [5], Media Richness Theory, which states that different media varies in the amount of information they allow to transit [6], and Social Identity Theory, which states that people establish a social identity as part of their self-concept by classifying themselves into specific social groups [7]. See Kaplan and Haenlein and Weinberg and Pehlivan [3,8] for two different social media classification frameworks.

Food blogs have become a popular platform for individuals to write about their recipes, restaurant meals, opinions, and food experiences in a public forum [9]. Everybody with a computer or a smart phone can post blogs, and for those who become popular enough to gain followers the pay may be very good. Users' present thoughts, feelings, likes and dislikes consistent with the

image they would like to give, and conveys product knowledge important for both brand and product managers. Some write very personal blogs where they invite the readers into their 'perfect' lives. Others write more dairy like blogs where they frequently post about their meals, wines and food tourism. These blogs can be interesting sources of information for food consumer research. By investigating blogs we can gain insights into how consumers shape and share their food identity [3] and how national food cuisines and local food practices can be transmitted to a global audience [10]. Microblogs like Twitter with spatiotemporal tagged information provide also an ideal source of data for investigating food exposure in real time [11]. Investigating blogs and bloggers behavior may display interesting knowledge and open up for new research questions. How does for example pre occupation with taking pictures of the food influence bloggers satisfaction with the meal?

Today food companies use social networking sites as Facebook to support the creation of brand communities, for conducting marketing research and even for distribution of food products. They use content communities as You Tube for involving customers in product related competitions. They ask customers to upload videos where they use the product, sing about the product or in other way interact with the product. Sometimes these competitions are too successfully, as in the 'Whopper Sacrifice' campaign where the fast food giant developed a Facebook application that gave users free Whopper sandwich for every 10 friends they deleted from their Facebook network. The campaign was adopted by over 20 000 uses, resulting in the sacrificing of 233 906 friends in exchange for free burgers before it was shut down by Facebook after one month [3].

Crowdsourcing

Social media made crowdsourcing, coined by Howe [12] as 'the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call', possible. Firms apply crowdsourcing for monitoring customers' interest, for gathering new ideas, and for creation of new products [13••]. PepsiCo monitored for example thousands of conversations with customers on social media and assessed customers' preferences via Facebook when formulating the new product Gatorade [8]. Danone encouraged consumers to participate in the creation of new cream desserts flavors, and when Lay's invited consumers to come up with a new potato chip flavor they got 245 825 proposals. After screening all the ideas a jury picked two winning flavors that where launched in the market. The two creator finalist were then endeavored to convince consumers to vote for their respective flavor. In addition to seeing her name on the product, the winner received €25 000 and 1% of the product's sales for a year [14•].

Studies investigating if users really can compete with professionals in generating new product ideas have found that on average user ideas score higher in novelty and customer benefit, but lower on feasibility [15,16••]. Experts seem to generate ideas that are easy to implement, while users generate ideas with a larger market potential. These findings support the importance of crowdsourcing for new ideas. When investigating social media-enabled customer co-creation projects in Barilla, Martini, Massa and Testa [17•] found that customers had a tendency to propose nostalgic products. The leading Italian pasta company created Facebook pages for each of its main products and organized an online community for all their brand lovers. They wanted to create a communication and relationship platform to give all people a voice; a platform where customers could submit ideas that others could vote for. The first year of activity, 4120 ideas were recorded. Based on votes two ideas were implemented. The community seemed to be driven by people who strived to re-create the past by proposing re-editions of old products, old packaging and old gadgets. Most of the ideas were also exploitative by nature, meaning not very radical.

While crowdsourcing seem to generate a win–win situation by creating value for both the firm and the customers, some obstacles have been found. Investigations of five FMCG case studies revealed two negative consumer reactions to crowdsourcing practices; the feeling of not being rewarded for their effort and the feeling of being cheated [14•]. The feeling of being cheated was essentially linked to the complexity of the crowdsourcing operation rules and customers misunderstanding of these rules. These concerns need to be taken into consideration in future crowdsourcing projects. Another worry that makes some firms reluctant to crowdsource is the necessary speed of the development project. To create, pack and brand a product in almost no time, which is necessary to keep the crowd's interest, means to rush the normal development process. Not all firms are willing to do so [14•]. Crowdsourcing may also influence the employees, who might worry about job security [13••].

Communication

Social media opens up for new ways of communicating with food consumers about health and food safety issues. Research indicate that between 55% and 67% of US adults go online for health and wellness information, and that social media is used by 20–34% of those searching for health related topics. Social media can facilitate social support groups, deliver education programs, recruit for services, train students, and help with communication between other health care professionals [18]. The web offers numerous advantages for those seeking health information, including anonymity, privacy, tailored health information, and the potential for interactivity and social support. Many researchers investigate how to influence healthy behavior via social media right

now. To mention a few: firstly, McKinley and Wright [19] show that social support leads to more online information seeking and a more favorable impressions of the health information on the web; secondly, Bottorff, Struik, Bissell, Graham, Stevens, Richardson [20] found that youth where positive to share YouTube videos designed to raise awareness among adolescent about tobacco exposure; and thirdly, Strarositu and Kim [21] show that low number of shares and comments on a cancer story posted on Facebook made people perceive the content to be stronger for others than for them self.

European Food Safety Association's (EFSA's) advisory group on risk communication recommends in their last report social media as a useful tool for disseminating food safety information [22]. They state social networking sites as Facebook to be good for rapidly informing a broad range of consumers about a simple message, blogs to be good for sharing reflective opinion pieces and microblogs as Twitter to be good for sending fast, topic-related alerts and for driving subscribers to online content where there is more information. Although the interactive aspect of social media offer the potential of effective food risk and crisis communication (via engagement and feedback), the road is not free of pitfalls. Social media makes it easy for anyone to put information on the internet, which can lead to credibility concerns [23]. While stakeholders and food safety experts across European view the opportunities that social media offers as higher than the threats, they are afraid that social media may escalate a food crises situation and create potentially unwarranted panic and hysteria [24•]. How social media affects consumers' food risk perception under crises is a topic that deserves further investigations.

Challenges with big data

The large amounts of data created by the new digital technologies can be hard to handle [25]. This data, often referred to as 'big data', comes from several different sources and varies in terms of *volume*, *variety* and *velocity*. Together, these three components determines the 'Bigness' of the data [26]. Big data can be small in terms of volume and velocity, but may come in an unstructured non-tabular fashion. This requires tailor made computer code in order to format the data. If data is small in terms of velocity and variety, but big in volume, it may require computational power far greater that any standard laptop. Finally, if data is small on variety and volume, but big on velocity, it requires continuous computations in real time, in order to utilize the full potential of the data [27].

Despite the claim that both existing sampling methods and the statistical tools associated with traditional consumer and sensory studies may be perfectly suited for answering a broad range of current and future research questions, they do not scale well when it comes to utilizing the potential of big data. Collecting big data requires programming and database management skills. Only rarely are the data in a format that is ready for analysis. Therefore the data requires additional processing. When dealing with big data, traditional *t*-tests and p-values becomes completely redundant, because the population size is often so large that everything becomes significant. Rather than relying on regressions and cluster analysis, big data analysis relies on machine learning techniques, which includes artificial neural networks, naïve bayes classifiers, K-nearest neighbors' classifiers, support vector machines and decision trees etc. [28]. All of them have their advantages and disadvantages, when it comes to interpretability, computational requirements, accuracy and the tendency to over fit.

Conclusion

Social media offers clearly many opportunities for sensory and consumer science. Utilizing the wisdom of the crowd can lead to a more user driven development process for food. By monitoring the conversations in social media we can generate new product ideas but also gain insights into what health and safety issues consumers are worried about. We can include our users into the development of new food products and make the crowd vote for the best solutions. The interactive aspect of social media opens up for new ways of communicating with the users. In spite of the many opportunities social media offers, there are also some obstacles that future research needs to address. We need research that investigates, firstly, how to handle the large amount of data generated by social media in an efficient way; secondly, how to hinder that social media escalates a food crises situation into unwarranted hysteria; and thirdly, how to utilize the crowd without making the contributors feeling exploited and cheated.

Acknowledgement

The research leading to this review study has been carried out within the project Orchestrating Food Innovation (Project # 225346) funded by the Agricultural Food Research Foundation of Norway.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Habibi MR, Laroche M, Richard M-O: **Brand communities based on social media: how unique are they? Evidence from two exemplary brand communities**. *Int J Inf Manage* 2014, **34**:123-132.

2. Gillin P: *The New Influencers: A Marketer's Guide to the New Social Media*. Sanger, CA: Quill Driver Books; 2007, .

3. Kaplan AM, Haenlein M: **Users of the world, unite! The challenges and opportunities of social media**. *Bus Horizons* 2010, **53**:59-68.

4. OECD: *Participative Web and User-Created Content: Web 2.0, Wikis, and Social Networking*. Paris: Organisation for Economic Co-operation and Development; 2007, .

5. Short J, Williams E, Christie B: *The Social Psychology of Telecommunications*. Hoboken, NJ: John Wiley & Sons, Ltd.; 1978, .

6. Daft RL, Lengel RH: **Organizational information requirements, media richness, and structural design**. *Manage Sci* 1986, **32**:554-571.

7. Tajfel H, Turner JC: **The social identity theory of intergroup behavior**. In *Psychology of Intergroup Relations.* Edited by Worchel S, Austin WG. Chicago: Nelson Hall; 1985:7-24.

8. Weinberg BD, Pehlivan E: **Social spending: managing the social media mix**. *Bus Horizons* 2011, **54**:275-282.

9. McGaughey K: **Food in binary: identity and interaction in two German food blogs**. *Cult Anal* 2010:69-98.

10. Lee SH, Samdanis M, Gkiousou S: **Hybridizing food cultures in computer-mediated environments: creativity and improvisation in Greek food blogs**. *Int J Hum–Comp Stud* 2014, **72**:224-238.

11. Chen X, Yang X: **Does food environment influence food choice? A geographical analysis through "tweets"?**. *Appl Geogr* 2014, **51**:82-89.

12. Howe J: **The rise of crowed sourcing**. *Wired* 2006, **14**:176-183.

13. •• Chiu C-M, Liang T-P, Turban E: **What can crowdsourcing do for decision support?** *Decis Supp Syst* 2014 http://dx.doi.org/10.1016/j.dss.2014.05.010.
This paper presents a framework of the roles of a crowd in different decision making phases.

14. • Djelassi S, Decoopmen I: **Customers' participation in product development through crowdsourcing: issues and implications**. *Ind Market Manage* 2013, **42**:683-692.
This paper gives good examples and explanations for the negative aspects of crowdsourcing.

15. Kristensson P, Gustafsson A, Archer T: **Harnessing the creative potential among users**. *J Prod Innov Manage* 2004, **21**:4-14.

16. •• Poetz MK, Schreier M: **The value of crowdsourcing: can users really compete with professionals in generating new product ideas?** *J Prod Innov Manage* 2012, **29**:245-256.
This paper presents the first real world comparison of ideas generated by a firms professionals with those generated by users in the course of an idea generation contest.

17. • Martini A, Massa S, Testa S: **Costumer co-creation projects and social media: the case of Barilla of Italy**. *Bus Horizons* 2014, **57**:425-434.
This paper gives a good insight into what food companies can expect as outcomes from crowdsourcing.

18. Tobey LN, Manore MM: **Social media and nutrition education: the food hero experience**. *J Nutr Educ Behav* 2014, **46**:128-133.

19. McKinley CJ, Wright PJ: **Informational social support and online health information seeking: examining the association between factors contributing to healthy eating behavior**. *Comp Hum Behav* 2014, **37**:107-116.

20. Bottorff JL, Struik LL, Bissell LJL, Graham R, Stevens J, Richardson CG: **A social media approach to inform youth about breast cancer and smoking: an exploratory descriptive study**. *Collegian* 2014, **21**:159-168.

21. Stavrositu CD, Kim J: **Social media metrics: third-person perceptions of health information**. *Comp Hum Behav* 2014, **35**:61-67.

22. EFSA: **When food is cooking up a storm**. *Proven Recipes Risk Commun* 2012:26-27.

23. Rutsaert P, Regan A, Pieniak Z, McConnon A, Moss A, Wall P, Verbeke W: **The use of social media in food risk and benefit communication**. *Trends Food Sci Technol* 2013, **30**:84-91.

24. • Rutsaert P, Pieniak Z, Regan A, McConnon A, Kuttschreuter M, Lozano N, Guzzon A, Santare D, Verbeke W: **Social media as a useful tool in food risk and benefit communication? A strategic orientation approach**. *Food Policy* 2014, **46**:84-93.
This paper gives a good overview of social media as a tool for risk and benefit communication.

25. Brynjolfsson E: **The four ways IT is driving innovation**. *MIT Sloan Manage Rev* 2010, **51**.

26. Gobble MM: **Resources: big data: the next big thing in innovation**. *Res Technol Manage* 2013, **56**:64-67.

27. Zaslavsky A, Perera C, Georgaskopoulos D: *Sensing as a Service and Big Data*. Bangalore, India: Presented at the International Conference on Advances in Cloud Computing (ACC-); 2012, .

28. Christensen K: *Mining the wisdom of the crowds: detecting new product ideas by text mining and machine learning techniques*. . (unpublished MSc thesis) Aarhus University, BSS; 2013.

Paper 3

Kasper Christensen, Kristian Hovde Liland, Knut Kvaal, Einar Risvik, Alessandra Biancolillo, Joachim Scholderer, Sladjana Nørskov and Tormod Næs

Mining online community data: The nature of ideas in online communities (Submitted for publication in Food Quality and Preference)

Mining online community data: The nature of ideas in online communities

Food Quality and Preference

Kasper Christensen[1,2,*], Kristian Hovde Liland[2,3], Knut Kvaal[1] , Einar Risvik[2] , Alessandra Biancolillo[2,4], Joachim Scholderer[5,6], Sladjana Nørskov[7], Tormod Næs[2,4]

[1] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway
[2] Nofima A/S, Ås, Norway
[3] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway
[4] Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Denmark
[5] CCRS and Department of Informatics, University of Zurich, Switzerland
[6] Department of Economics and Business Economics, Aarhus University, Denmark
[7] Department of Management, Aarhus University, Denmark
* Corresponding author. Email: kasper.christensen@nofima.no, Tel.: (+47) 94 15 89 93

Target journal: Food Quality and Preference

Word count: 11,158

Abstract

Ideas are essential for innovation and for the continuous renewal of a firm's product offerings. Previous research has argued that online communities contain such ideas. Therefore, online communities such as forums, Facebook groups, blogs etc. are potential gold mines for innovative ideas that can be used for boosting the innovation performance of the firm. However, the Big Data nature of online community data makes idea detection labor intensive. As an answer to this problem, research has shown that it might be possible to detect ideas from online communities, automatically. Research is however, yet to provide an answer to what is it that makes such automatic idea detection possible?

Our study is based on two datasets from dialogue between members of two distinct online communities. The first community is related to beer. The second is related to Lego. We generate machine learning classifiers based on Support Vector Machines and Partial Least Squares that can detect ideas from each respective online community. We use partial least squares to investigate what are the words and expressions that allows for automatic classification of ideas. We conclude that ideas from the two online communities, contains suggestion/solution words and expressions and it is these that make automatic idea detection possible. We conclude that the nature of the ideas in the beer community seems to be related to the brewing process. The nature of the ideas in the Lego community seems to be related to new products that consumers would want.

1. Introduction

Online communities can be important drivers of knowledge generation for the firm. They allow people with similar interests to gather and interact, even though these people are geographically far from one another. Thus, online communities become locus points for people all over the world, and they enable people to interact and unite their shared knowledge. This makes room for *new* knowledge generation that can be used to innovate the firm and *our* society on a continuous basis (von Hippel, 2001; Lee & Cole, 2003; Jeppesen & Frederiksen, 2006). Facebook groups, google forums and newsgroups are all examples of online community types that allow for knowledge sharing and knowledge generation related to a given topic.

A special kind of knowledge that has occupied innovation management scholars and R&D people is *ideas* (Kristensson, Gustafsson, & Archer, 2004; Dean, Hender, Rodgers, & Santanen, 2006; Magnusson, 2009; Magnusson, Wästlund, & Netz, 2014; van den Ende, Frederiksen, & Prencipe, 2015). Ideas represent a very specific kind of information and it has been claimed that ideas often contain both problem- and solution information related to a given topic (Poetz & Schreier, 2012; van den Ende et al., 2015). Therefore, firms are eager to secure a continuous stream of ideas. To achieve this goal, some firms have established their own online communities, where dedicated product users and consumers gather to discuss- and suggest ideas to the firm. For instance, the computer manufacturer Dell (di Gangi, Wasko, & Hooker, 2010) and the music software manufacturer Propellerhead (Jeppesen & Frederiksen, 2006) both rely on user idea generation stemming from their own online communities for improving the firms innovation performance and enhancing a long term competitive advantage.

The online communities associated with Dell and Propellerhead are *firm*-hosted communities, because they are hosted by the firm itself. However, online communities do not need to depend on a firm to provide an ideation platform and maintain activities. Thus, another widespread type of community exists, namely the type that is established by the users of the community itself. This type of community exists *independently* of a firm and this "firm-free" online community is self-supporting, self-sustaining and it is typically centred on products, activities or brands (Füller, Jawecki, & Mühlbacher, 2007). Examples of firm-free communities that have been studied involve communities focusing on sailplaning, canoeing, bordercross, cycling, basketball and Lego (Franke & Shah, 2003; Antorini, Muñiz, & Askildsen, 2012)

As opposed to the firm-hosted communities related by Dell and Propellerhead, the free online communities are *not* based on software designed to enable harvesting of ideas and knowledge generated by the community. This implies that if a researcher or a firm wants to benefit from the ideas and the knowledge generated by the free online community, the only existing solution is to read everything written and to filter the relevant information *manually*. Manual filtration is by-en-large unfeasible, as the information stored in each community accumulates into several thousand- if not millions of text pieces that have been exchanged between community members over time (Lin, Hsieh, & Chuang, 2009)

In an attempt to handle this filtration problem, it has been demonstrated that ideas from a free online community related to the product Lego, can to some extent, be automatically identified and extracted via a type of artificial intelligence system, relying on text mining and machine learning. The system takes as input a lot of idea texts and *non*-idea texts and in this way, the system learns what characterizes idea texts in opposite to non-idea texts (Christensen, Nørskov, Frederiksen, & Scholderer, 2016). The described system is based on a machine learning technique named *Support Vector Machines*. Support vector machines

are known for their high and robust performance on high dimensional sparse datasets and are therefore a good choice for text classification problems. The downside of using support vector machines is that they are *un*-transparent (Kotsiantis, Zaharakis, & Pintelas, 2007), meaning that it is not easy to understand and explain how classifications are made when utilizing this particular machine learning technique.

The lack of transparency is usually not a problem when using support vector machines *purely* for classification and prediction. As long as the support vector machine is thoroughly tested with consistent and robust results, it should in theory be able to find what it is supposed to find. However, the lack of transparency becomes a problem when we, the users of the method, *also* seek to explain the underlying phenomenon that enables automatic classification. And, if future research want to aim at improving data representation and methodology on this kind of text classification problem, it is important that future methods are designed in a way that gives insights into relations that drives classification. That is the point of departure for this paper.

The present paper has two scopes: First of all it is an investigation of whether a well-known and much used method in the area of sensory- and consumer science, *Partial Least Squares* (see e.g. S. Wold, Martens, & Wold, 1983; Martens & Næs, 1991) can provide the additional interpretation power that the support vector machine lacks. Partial least squares regression is a method that has proven to be very useful for classification as well as for interpretation of the relations that drives classification. Therefore partial least squares might provide us with insights on what are the words and expressions that are driving automatic classification of ideas written in online communities. What is the nature of ideas written in online communities? An integral part of this investigation will be whether the partial least squares

technique is comparable to the support vector machines when it comes to classification power.

The second scope is to extend the approach used in (Christensen et al., 2016) to also take into account *doubt* texts, i.e. texts which are not easily classified as either an idea text or non-idea text. This is a highly relevant situation in practice, which in this case will be achieved by incorporating an extra class in the testing of the classifiers representing texts in which *also* the classifiers were in doubt. Two and very different online communities cases, Lego and beer brewing, will be used for evaluating the methodology.

The paper is structured so that in the next section we introduce the class of methods we have chosen. Here, we introduce the different text mining- and machine learning concepts from a general point of view. Next, we introduce the classification techniques, support vector machines and partial least squares and we describe how to use them for text classification and how to use partial least squares for variable selection and interpretation. Then, we report our exact method for collecting and analysing data. We report our results, and we end the paper with a discussion and a conclusion.

2. Choice of methods

Our supervised machine learning procedure for idea detection can be divided into four main parts (See Figure 1 for overview).

2.1 - Data collection and target generation

The first part is to identify a data source of interest, extract the texts from the same data source and generate a target variable. To generate the target variable, human raters can be introduced, and *how* the target variable is generated is especially important, since it is the target variable that contains the information the machine learning technique uses for learning. In order to generate a target variable for text classification tasks, crowdsourcing can be used (Howe, 2006; Wang, Hoang, & Kan, 2013; Christensen et al., 2016). When utilizing crowdsourcing for this type of task, Sautter & Böhm (2013) argue that two main sources of error exists. The first source is *honest misjudgments*. This type of error is related to the likelihood of the raters (also called crowdworkers or workers) making misjudgments if the crowdsourcing task is complex. The second source is *dishonest crowdworkers* and this source of error is related to crowdworkers who are not doing the work they are supposed to do for opportunistic reasons (Eickhoff & de Vries, 2013; Wang et al., 2013). The presence of these potential error types means that if no precautions are taken, the evaluations made by the raters can become inflated by erroneous responses. Therefore it is desired to apply strategies that ensure high quality ratings.

When the text data has been extracted and the target variable has been generated, the dataset can be presented as shown in Table 1. In this table the "Id" column takes a unique value for each text. The "Target" column takes on the two values 0 and 1 depending on whether the text is considered within or outside of the actual criterion set (i.e. idea text vs. non-idea text). The "Text" column contains the actual text written in the online community.

Table 1 - Imaginary examples online community texts coupled with a target variable and an id variable

Id	Target	Text
id001	0	´My wife is having a party for all of her friends and she asked me to make a nice beer for them to drink. What should I make them?´
id002	1	´You can make a strawberry beer. I made one for my wife's birth day and it was a huge success!´
id003	0	´I don't know what do to about this problem. It has been bugging me for weeks and I cannot find a solution to this! Someone please help me!´

2.2 - Text pre-processing and partitioning

In text mining, the pre-processing step is where the raw text is turned into a numerical format that can be used for machine learning. In this process, all punctuation marks are removed. All numbers are removed. All upper case letters are converted to lower case letters, all extra whitespaces are removed and usually all stopwords are removed. Stopwords are words that are so common in our written language that they are assumed to have no discriminative power (Feinerer, Hornik, & Meyer, 2008). These steps are standard text cleaning steps. Finally, all remaining words are counted with respect to how many times they occur in each and every text. The output of this procedure is known as the *bag-of-words* (Linoff & Berry, 2011).

One problem with the bag-of-words concept is that in its most simple variant relations between- and the order of the words is lost. In order to compensate for this problem, n-grams can be generated. N-grams refer to *series* of words. For example "let me think" is a trigram, "good idea" is a bigram and "beer" is a single word, but can also be defined as a unigram. N-grams are useful because they allow the words to interact and by applying n-grams, more information is kept in the dataset (Zanasi, 2007; Radovanović & Ivanović, 2008) (From now on we refer to single words and n-grams by using the word *terms*). The generation of the bag-

of-words will often result in a large dataset, that has as many rows as there are texts in the raw dataset, and as many columns as there are unique terms in that same dataset. 1,000, 10,000 or 100,000 terms are not uncommon. Therefore, it can be an advantage to reduce the column dimension of the bag-of-words by removing *sparse terms*. Sparse terms refer to terms that are *extremely* rare. And, because they are rare, the assumption can be made that they are *so* specific that they do not hold any predictive power. The removal of these terms will have the effect that redundant terms are filtered and computation time is reduced (Tirunillai & Tellis, 2014; Antons, Kleer, & Salge, 2016; Christensen et al., 2016). Sparse terms can be removed by setting a sparsity threshold. For example a sparsity level equal to 1% means that, terms that are in less than 1 out of 100 of the texts will be ***ex***cluded. The result of the text pre-processing is shown in Table 2. Here the "Id" and the "Target" remain the same as in Table 1, *but* the "Text" column has been converted into columns representing the counts of distinct terms (See Feldman & Sanger (2006) and Feinerer, Hornik, & Meyer (2008) for more details on text pre-processing). When the bag-of-words has been generated, it should be partitioned into a training set, a validation set and a hold-out set.

Table 2 - Example of bag-of-words

Id	Target	Term 1	Term 2	Term3	Term 4	Term n
id001	0	1	0	1	1	0
id002	1	0	1	0	0	0
id003	0	0	0	0	1	1

2.3 - Training and testing

The training set and validation set are used for training and tuning the machine learning classifier (Tuning is sometimes called calibration). The hold-out set is used for assessing over-fit and over-generalization, and it is the only dataset that is valid when assessing (true) performance of the trained classifier (Hastie, Tibshirani, & Friedman, 2008).

In the validation process the classification power of the classifier is obtained with respect to predefined performance measures. Several performance measures can be used and some of the most common ones are precision (specificity), recall (sensitivity) and the so-called F_1-measure (Witten & Frank, 2005):

$$\text{Precision} = \frac{\text{True positives (TP)}}{\text{True positives (TP)} + \text{False positives(FP)}} \quad (1)$$

$$\text{Recall} = \frac{\text{True positives (TP)}}{\text{True positives (TP)} + \text{False negatives (FN)}} \quad (2)$$

$$F_1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

The performance measures are calculated based on counts of how many texts the classifiers correctly classifies. In our case a true positive (TP) is a text that has been classified as idea text by the raters and the classifier. A true negative (TN) is a text that has been classified as a *non-idea* text by the raters and the classifier. A false positive (FP) is a text that has been classified as idea text by the classifier but *not* the raters. And, a false negative (FN) is a text that has been classified as non-idea text by the classifier but not the raters. Recall can be interpreted as the proportion of true positives that the classifier identified. Precision can be interpreted as the proportion of identified positives that are in fact *true* positives. F_1 seeks to balance precision and recall.

Before the counts can be obtained and performance can be calculated, a cut-off threshold will have to be set. Most classification techniques can give as output real numbers (i.e. 0.25, 0.67 or 0.97) instead of binary codes (i.e. 0 or 1). It is useful to manually set the cut-off threshold because a well-known problem related to classification tasks can be handled. This problem is known as the *class imbalance* problem and it is related to the scenario where the class distribution in the dataset is skewed (i.e. there are more texts belonging to the non-idea class than the idea class, or vice versa). If not handled, this problem can lead to

classification skewness towards the majority class, meaning that the classifier can potentially end up classifying almost all texts as belonging to the majority class. There are several strategies for handling this problem and it is not always clear what strategy is the best, because the best strategy is often dependent on the data (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). Christensen et al. (2016) used a strategy known as *under-bagging*, but Provost (2000) suggests that sometimes the simplest solution is to set the cut-off threshold so that performance is optimized. For example if a classifier has assigned a classification score 0.75 to text one and 0.60 to text two, the cut-off threshold can be set to 0.65 so that text one gets binary code 1 and text two gets binary code 0. By doing this for all classification scores in the validation set, a vector of *predicted* target values is obtained. This vector can be compared to the vector of *true* target values and in this way the cut-off threshold can be adjusted so that the classifier correctly classify as many texts as possible in the validation set.

Next, the classifier is used for classification on the hold-out set and performance obtained on the validation set is compared with the performance obtained on the hold-out set. Based on this comparison, classification performance, over-fit and over-generalization can be assessed. Over-fit refers to the case where a model obtains high performance on the validation dataset but low performance on the hold-out set. In practice this means that the classifier has been trained to find texts that look almost *exactly* like the training- and the validation texts. This is not preferable because the classifier would most likely never occur texts that looks exactly like the training texts (Hastie et al., 2008; Linoff & Berry, 2011). Therefore, it is preferable to train a classifier to find *approximately* (generally) the same pattern represented in the training texts. On the other hand side, over-generalization (also called under fitting) refers to the case where a classifier performs low on both the validation set and the hold-out set. This is obviously also not preferable.

A final aspect that we suggest is taken into account is *doubt* texts. This aspect stems from the observation that the raters in Christensen et al. (2016) *disagreed* with respect to 197 texts out of 3,000 texts. These 197 doubt texts texts were omitted from further analysis. This aspect is important in our case because we do not expect that the raters can obtain perfect agreement when generating the target variable. And further, if the raters cannot agree, it is not reasonable to expect the classifiers to do a perfect classification either. By setting different cut-off thresholds, one can incorporate the aspect of *doubt* into the testing of the classifier *instead* of omitting the data for testing. In this way it can be investigated if the classifier displays the same degree of doubt as did the raters when *they* classified the texts. For example one can define *three* classes, idea texts, non-idea texts and doubt texts. This means that *two* cut-off thresholds will have to be set instead of one: An upper threshold for the *clear* idea texts and a lower threshold for the *clear* non-idea texts. Doubt texts are defined as all texts where the classification scores falls *in between* the upper- and the lower cut-off threshold.

2.4 - Interpretation

Based on the performance measures, the classification performance of the classifier can be determined and in theory, interpretation of the results can be performed by proper visualisation techniques and methods for interpretation. However, when dealing with many terms or variables, literature suggests that the set of terms used for generating the classifier should be reduced. There are two arguments for this. The first argument is that some classification techniques do not cope well with *too* many terms and can become computational expensive/impossible and loose predictive power (Forman, 2003). But more importantly, the abundance of terms makes interpretation difficult if not impossible (Guyon & Elisseeff, 2003). This is supported by Mehmood, Liland, Snipen, & Sæbø (2012) who write that one of the main motivations for performing term/variable selection is to identify a set of *highly* important terms, thus allowing for easier interpretation of the results. This set of terms

can then serve as focus for further analysis. And further, another similar view on this matter can be found in Martens (2001) who writes that especially the partial least squares technique gives *cognitive access* to the relevant information in the data. Here, the concept of cognitive access refers to *our* ability to understand the data. This viewpoint is central because it is the reduction in terms coupled with a proper strategy for interpretation that gives us cognitive access. Thus, to achieve a good starting point for interpretation a two step-process is appropriate. In the first step a computational procedure is used to reduce the set of terms so that only the most predictive terms remain. These terms can then be visualized and a semantic approach can be used to interpret the meaning of the most predictive terms.

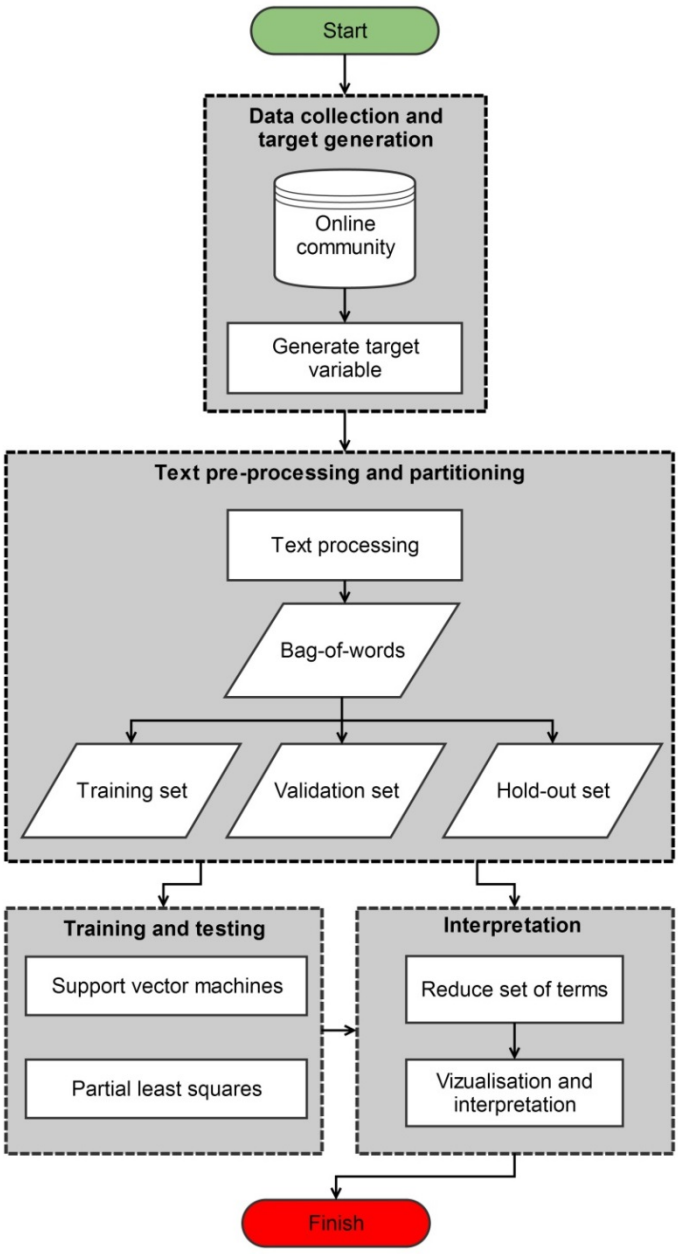


Figure 1 - Flowchart showing the supervised machine learning procedure for classification and interpretation that we apply.

3. Classification techniques

3.1 - Classifier selection

Many techniques exists that can be used for generating classifiers for different purposes. For example Caruana & Niculescu-Mizil (2006) did a comparison of 10 algorithms: Support Vector Machines, Neural Nets, Logistic Regression, Naïve Bayes, Memory Based Learning, Random Forests, Decision Trees, Bagged Trees, Boosted Trees and Boosted Stumps. In a similar manner, Amancio et al. (2014) did a comparison of Naïve Bayes, Bayesian Networks, different versions of Decision Trees, K-Nearest Neighbors, Logistic, Multilayer Perceptrons and Support Vector Machines. Many studies of this type exist, and often the focus of such studies is on the predictive power of the algorithms. However, only few studies focus on the transparency of the different techniques and a study that *did* focus on transparency is Kotsiantis et al. (2007). Here, the authors did a qualitative comparison of machine learning classifiers and concluded that the support vector machine is one of the most accurate algorithms (high predictive power), but also one of the least transparent.

3.2 - Support vector machines

Support vector machines were developed in the 90'ies and can in its most simple form, be considered a discriminant analysis that allows for data structures that are not linearly separable (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995). Support vector machines utilise a complex transformation of the original data space, into a higher order space where the data points can be more easily separated. This transformation is known as the *kernel trick,* and it is this specific trait that makes the support vector machine non-transparent, *but* also capable of handling the many predictor variables (i.e. terms) and non-linear relations. Support vector machines, come in several variants, where the *linear* support vector machine is probably the most common.

Mathematically, a linear support vector machine generates a decision boundary that best separates two classes (Idea texts from non-idea texts). Also, it generates a *margin* surrounding the decision boundary. The soft margin constant *C* controls the position and width of the margin and can be seen as a parameter that is used to adjust the classifiers sensitivity to misclassifications. Thus, the smaller C is, the wider the margin becomes and the higher is the tolerance to misclassifications. In practice C is found computationally, by defining a range of C values and training one classifier for each C value. Next, the series of classifiers are tested and the best C value is the one associated with the highest performing classifier (See Cortes & Vapnik (1995) and Ben-Hur & Weston (2010) for further details on support vector machines)

3.3 - Partial least squares

Another powerful technique that can be used for machine learning and classification is *Partial Least Squares*, also known as PLS. The partial least squares technique has primarily been used in chemometrics and sensometrics, but also in social sciences because it is capable of handling large datasets with many predictor variables and few observations (Wold, Sjöström, & Eriksson, 2001; Næs, Isaksson, Fearn, & Davies, 2002). Partial least squares can be used for both regression and classification (Wold et al., 1983; Martens & Næs, 1991). When used for classification, a regression problem is transformed into a classification problem by introducing a "dummy" matrix consisting of 1's and 0's, signaling the idea class and non-idea class respectively.

When used for text mining, partial least squares compresses the set of terms into a reduced set of components, or latent variables, by maximizing the covariance between the terms and the target. Here, the texts are projected into a lower dimensional subspace, and then the latent variables are used *instead* of the original terms to estimate the target value. The component-wise modelling of the connection between terms and target, results in a series of

text coordinates (scores) and term coordinates (loadings) which can be plotted and/or interpreted. It is these specific traits that make partial least squares more suitable for interpretation than support vector machines, because we, the user of the technique, do *not* lose the information about the terms as we do with support vector machines.

To reduce the set of terms, many different variable/term selection methods can be used with partial least squares, and they use different model parameters or a priori knowledge to evaluate the relevance of the terms. In this work, we have focused our attention on one specific variable selection method called *Significance Multivariate Correlation* (Tran, Afanador, Buydens, & Blanchet, 2014). Significance multivariate correlation is a signal-to-noise measure calculated as the ratio between the variance explained from each term and its residual. To compute significance multivariate correlation from partial least squares, the entire set of terms is used to create a partial least squares classifier. Then, significance multivariate correlation values are estimated and the terms are systematically/iteratively reduced to a smaller submatrix where the most predictive terms remain for interpretation. Further details on the partial least squares algorithm can be found in (Ståhle & Wold, 1987; Barker & Rayens, 2003).

4. Materials and methods

4.1 - Data collection and target generation

Our two online communities are of the *Usenet Newsgroup* type. Usenet is an internet service that hosts a gigantic network of newsgroups. It was established in the 1980'es. Everyone can sign-up for and/or host own newsgroups for a small amount of money[1]. When participants sign up, they can write posts and comments to other members of the newsgroup and discussion and ideation related to the specific topic of the newsgroup is generated.

Our beer case is a home brewing community. In this community participants discuss topics related to home brewing, and we expected that interesting ideas related to beer could be found in this community. The community contains 10,528 texts in total and the first text was written in June 2003. The last text was written in July 2014. Our second case is a Lego community, where participants discuss- and suggests ideas to new Lego products. The community contains 5,652 texts and the first text we can register was written in September 1998 and the last text was written in October 2012. Both communities are firm free in the sense that no firm has initiated the community and no firm has control over the community. The data from both communities are freely accessible via the internet.

We used crowdsourcing for generating the target variable. To ensure high quality evaluation by the crowd workers, we introduced training texts. Examples of the training texts are displayed further below in Table 4.1 and Table 4.2. For a crowdworker to qualify for the idea rating task, the crowdworker should answer correctly 4 out of 5 training texts. Elsewise the crowdworker would not be allowed to enter the task. If a crowdworker was allowed to enter the task, the crowdworker would have to evaluate the texts in series of five texts at the time. If the crowdworker did not maintain 80% accuracy throughout the whole task, the

[1] http://www.usenet.net/usenet-faq/#_usenet

crowdworker was removed from the task with all of the ratings the crowdworker had made. Also, as soon as a crowdworker had been exposed to all training texts one time, the crowdworker would not be allowed to evaluate more texts.

We used five crowdworkers per text to create the target variable. The crowdworkers were recruited via Crowdflower, a crowdsourcing service. We defined a text as an idea text, if 5 out of 5 crowdworkers had agreed that the text contains at least one idea. Similarly, we defined a text as a non-idea text if 5 out of 5 crowdworkers had agreed that the text did not contain at least one idea. The rest of the texts were defined as *doubt* texts. The entire process resulted in two machine learning datasets: One for beer and one for Lego.

Table 4.1 - Two examples of crowd training texts from brewing community

Idea	**Non-idea**
´I have been thinking about formulating a November brew with a substantial (if not entirely) corn component since I usually have access to a lot of deer corn around that time. I hadn't thought about the oil aspect though. (November is the start of deer season in Texas). I thought it would be cool to have some deer corn beer at the Hunting lodge. This reminds me of a suggestion I saw on my packet of Koji, that it could be used on pearled barley or other grains to convert the starches. I suppose it could be used entirely on corn or even potatoes, but I wonder if anyone has tried it? Do you think it would have any effect on the oils?´	´I have a newer haier kegerator and my problem is that it does not seem to get cold enough even with the thermostat turned all the way to the cold position. The thing is that the inside metal cooling sheet in the back is icy so it seems like it running correctly. Since the entire sheet is ice I don't know how it could perform any better. Does anyone have any ideas what the problem could be. The specs for the model say it should get down to 38o but at its coldest setting my thermometer says 48o. This may be more of a refrigeration question then beer question but I couldn't find any refrigeration news groups. Thanks for any help.´

Table 4.2 - Two examples of crowd training texts from Lego community

Idea	Non-idea
´Dear Lego, Please consider making the Extra-Large baseplate (#628) in other colors. Colors which immediately come to mind are... Blue, Tan, White, Black and Brown. This could be a one time special production run, packaged one of each color as a set of 5 for $50 (the marketplace and hosetrading sorting out who really wants which color). Available only at S@H and possibly some of the LIC stores.´	´So I got the newest catalog tonight, and flipped through it . . . And suddenly, What the [lenghty diatribe, of the flavor that got Marchetti banned???!!!!!????!??! Okay, I realize that there's been inflation, and it's been maybe a decade since the last price increase, but $15 for a 48x48 baseplate? That's MdStone to boot? A 50% markup overnight? Arrghh.´

4.2 - Text pre-processing and partitioning

For both individual datasets, the texts were processed by removing all punctuation marks, removing all numbers as well, and removing all extra whitespaces. We generated all possible combinations uni-grams, bi-grams and tri-grams and we removed stopwords *after* the generation of n-grams. All unique terms were counted and the bag-of-words was pruned at a 0.20% level.

We excluded the doubts texts and the remaining texts were partitioned so that the training sets consisted of 70% of the texts, the validation sets consisted of 15% of the texts and the hold-out set consisted of 15% of the texts. We maintained the natural class balance in each partition. For the beer dataset, 405 of the texts were evaluated as idea texts and 988 texts were evaluated as non-idea texts. For the Lego dataset, 515 texts were evaluated as idea texts and 798 texts were evaluated as non-idea texts. The bag-of-words contained 11.049 terms for the beer datasets and 12.426 unique terms for the Lego datasets.

4.3 - Training, tuning and testing

Four classifiers were trained, tuned and tested. That is: One support vector machine classifier for the beer dataset, and one for the Lego dataset. One partial least squares classifier

for the beer dataset and one for the Lego dataset. The training and tuning was done with respect to the F_1-measure. For the support vector machine classifiers, we used a linear kernel and the classifiers were tuned with respect to the soft margin constant C and the optimal cut-off threshold. The partial least squares classifiers were tuned with respect to number of components. Also the partial least squares classifiers were tuned with respect to the optimal cut-off threshold. For performance evaluation, we fixed the tuning parameters for the four classifiers and used the classifiers for classification on the hold-out sets. For assessment of over-fit and over-generalization, we compared performance of the classifiers obtained on the hold-out sets with the performance obtained on the validation sets (R Core Team, 2016).

4.3.1 - Classification performance on doubt sets

In order to incorporate the aspect of *doubt texts* into the modelling scheme, we applied the trained classifiers on the doubt text. Thus, we trained and tuned the classifiers by using only very clear idea texts and very clear non-idea texts. And, to investigate how the classifiers would perform *also* on the doubt texts, we separated the target variable in the doubt sets and the hold-out sets into six groups based on the crowdworkers ratings. If five out of five crowdworkers have evaluated a text as an idea text, the text is a *clear* idea text. If four out of five crowdworkers have evaluated the text as an idea text, the text is an *almost* idea text. If three out of five crowdworkers have evaluated the text as an idea text, the text is a *doubtful* idea text. If two out of five crowdworkers have evaluated the text as an idea text, the text is a *doubtful non*-idea text. If one out of five crowdworkers have evaluated the text as an idea text, the text is an *almost* non-idea text. If zero out of five crowdworkers have evaluated the text as an idea text, the text is a clear non-idea text.

Next, we defined two new cut-off thresholds for the classifiers. We defined a *clear* idea text as all texts that obtains classificaiton scores higher than or equal to 0.80. All *doubt* texts were defined as all texts that would score between 0.80 and 0.20. All clear *non-idea*

texts were defined as text that would score lower than or equal to 0.20. This resulted in a three-by-six matrix that would allow us to investigate if the classifiers show the same degree of uncertainty as did the crowdworkers when they classified the texts.

4.3.2 - Classification performance when short- and long texts have been removed

Before progressing to term selection and interpretation we did a second additional test where we removed short texts and long texts from both the beer dataset and the Lego dataset. The aim was to investigate how both of the classification techniques would respond to datasets that was more homogenous with respect to the number of terms in the training texts. Our argument for this focus is that some non-idea texts are very short, and they contain almost no terms. One the other hand side, idea texts are often lengthy because it is difficult to express an idea without using at least a couple of sentences. By having an overflow of short non-idea texts and lengthy idea texts in the training datasets, we could in the most extreme case train and tune classifiers that have only been taught to distinguish shorts texts from long texts. This is not preferable and we wished to investigate to what degree this was the case.

Therefore, we excluded a dramatic amount of training texts based on visual inspection of the term distribution for the documents. All text that contains less than 25 terms and all texts that contains above 100 terms were excluded from the training texts. By doing this we had generated more homogenous training datasets with respect to the length of the texts. After removal of short and long texts, the beer dataset consisted of 503 texts (192 idea texts and 311 non-idea texts). The Lego dataset consisted of 473 texts (106 idea texts and 367 idea texts). The natural class balance was maintained. We partitioned the datasets and trained, tuned and tested the classifiers by using these reduced sets of texts. We compared the performance of the new classifiers with the performance of the classifiers trained on the *full* sets of clear idea texts and clear non-idea texts.

4.4 - Term selection and interpretation

For each online community dataset (beer and Lego) we used the significance multivariate correlation approach for term selection. Instead of using the full sets of texts (1,393 texts for beer and 1,313 texts for Lego), we used the sets of clear idea texts and clear-non-idea texts where short texts and long texts had been removed (503 texts for beer and 473 texts for Lego). Our argument for this choice is that we were interested in investigating what terms that distinguish between clear idea texts and clear non-idea texts of similar size. Here *no* partitioning was done and the validity of the term selection was based a 10-fold cross validation.

We experimented with several settings for how many terms that would be removed for each iteration of the term selection procedure. We chose an approach where we iteratively removed the 0.5% worst performing terms. Here, the number of components was equal to two. This means that first one partial least squares classifier was generated based on the full set of terms. Then, the 0.5% *least* predictive terms were removed based on the significance multivariate correlation values. Next, a new partial least squares classifier was trained by using the *reduced set* of terms. Again the 0.5% least predictive terms was removed. This process continued until 50 terms remained and based on the these terms scores plots and loading plots were produced for visualization and interpretation of what terms (loadings) that is separating idea texts from non-idea texts (scores).

5 - Results

5.1 - Beer community

5.1.1 - Beer classification task

The results from the classification task on the full beer dataset show that the support vector machine classifier performed slightly better than the partial least squares classifier (Table 5). This applies for the F_1-measure, the precision measure and the recall measure. The support vector machine classifier shows no tendency to over-fit over-generalize, and neither did the partial least squares classifier. We notice that the partial least squares classifier is relatively simple as it utilizes two components. We do not consider the performance between the two classification techniques remarkably different.

Table 5 - Results of classification task on the full beer dataset

Technique	Tuning	Partition	Precision	Recall	F_1
Support vector machine	C = 1e-05 Cut-off = 0.52	Validation	0.95	0.93	0.94
		Hold-out	0.92	0.93	0.93
Partial least squares	Components = 2 Cut-off = 0.25	Validation	0.93	0.93	0.93
		Hold-out	0.92	0.92	0.92

When we applied the support vector machine classifier on the beer doubt set and the beer hold-out set it became clear that the texts that the crowdworkers could not agree on, were equally difficult for the support vector machine to classify (Table 6). The same trend applies for the partial least squares classifier (Table 7). However, the partial least squares behaved different than the support vector machine classifier with respect to two aspects. The first aspect is that the partial least squares classifier classified over 75% (46 out of 60) of the *clear* idea texts as doubt texts. The second aspect is that it classified over double as many texts as doubt texts, compared to the support vector machine. Here the support vector machine

classified 393 texts as doubt texts and the partial least squares classifier classified 1,054 texts as doubt texts.

Table 6 - Results from classification task on the beer doubt set and hold-out set with the support vector machine classifier

Support vector machine		Classifications		
		Idea (672)	Non-idea (750)	Doubt (393)
Crowd evaluations	Clear idea (60)	83.33% (50)	3.33% (2)	13.33% (8)
	Almost idea (371)	70.08% (260)	10.78% (40)	19.41% (71)
	Doubtful idea (415)	46.27% (192)	25.78% (107)	27.95% (116)
	Doubtful non-idea (375)	28.80% (108)	43.47% (163)	27.73% (104)
	Almost non-idea (446)	13.23% (59)	67.71% (302)	19.06% (85)
	Clear non-idea (148)	1.35% (2)	93.24% (138)	5.41% (8)

Table 7 - Results from classification task on the beer doubt set and the hold-out set with the partial least squares classifier

Partial least squares		Classifications		
		Idea (61)	Non-idea (700)	Doubt (1.054)
Crowd evaluations	Clear idea (60)	20% (12)	3.33% (2)	76.67% (46)
	Almost idea (371)	7.01% (26)	9.97% (37)	83.02% (308)
	Doubtful idea (415)	3.61% (15)	23.37% (97)	73.01% (303)
	Doubtful non-idea (375)	1.33% (5)	37.87% (142)	60.8% (228)
	Almost non-idea (446)	0.67% (3)	63.9% (285)	35.43% (158)
	Clear non-idea (148)	0.00% (0)	92.57% (137)	7.43% (11)

When short- and long texts were removed from the beer dataset, performance of the support vector machine classifier decreased in comparison to when the full set of texts was used (Table 8). Also the performance of the partial least squares classifier decreased. There is no noteworthy difference in the performance between the support vector machine classifier and the partial least squares classifier. However the results *do* suggest that both classifiers are victims to over-fitting to the training- and validation data since there is a relatively large difference in performance on the validation set and performance on the hold-out set. The difference in the results obtained on the *full* set of texts and this *reduced* set of text, suggest that the relatively short beer text and relatively long beer texts, are easier to separate for both of the classification techniques than are texts of similar length.

Table 8 - Results of classification task on the reduced beer dataset where short texts and long texts have been removed

Technique	Tuning	Partition	Precision	Recall	F_1
Support vector machine	C = 1e-06 Cut-off = 0.42	Validation	0.93	0.90	0.91
		Hold-out	0.73	0.76	0.75
Partial least squares	Components = 3 Cut-off = 0.31	Validation	0.96	0.86	0.91
		Hold-out	0.69	0.76	0.72

5.1.2 - Beer idea term selection and interpretation

We progressed with term selection with the partial least squares classifier on the reduced set of texts. Here, the significance multivariate correlation term selection procedure resulted in the selection of 50 terms. The scores plots (Figure 2) show a trend that suggests that non-idea texts and idea texts are separated by the two components but component one plays a larger role than does component two. Principal components analysis was used for the same data set and the scores plot turned out to be similar to the partial least square plot, supporting these findings. The loadings plot (Figure 3), show that terms like “if you”, “solution”, “you want”, “you can”, “thinking”, “the beer”, “sugar”, “flavor” and “yeast” are terms with high loadings for beer community ideas. The texts in Table 9 and Table 10 are examples of idea texts and non-idea texts. The text in Table 9 is an example of a typical idea from the beer community related to “how to do something” in the brewing process. To us, it makes sense that the text is an idea because it reflects suggestion dialogue from one community member to another. In Table 10, a non-idea text is shown, and to us this particular text can be regarded as a comment from one community member to another.

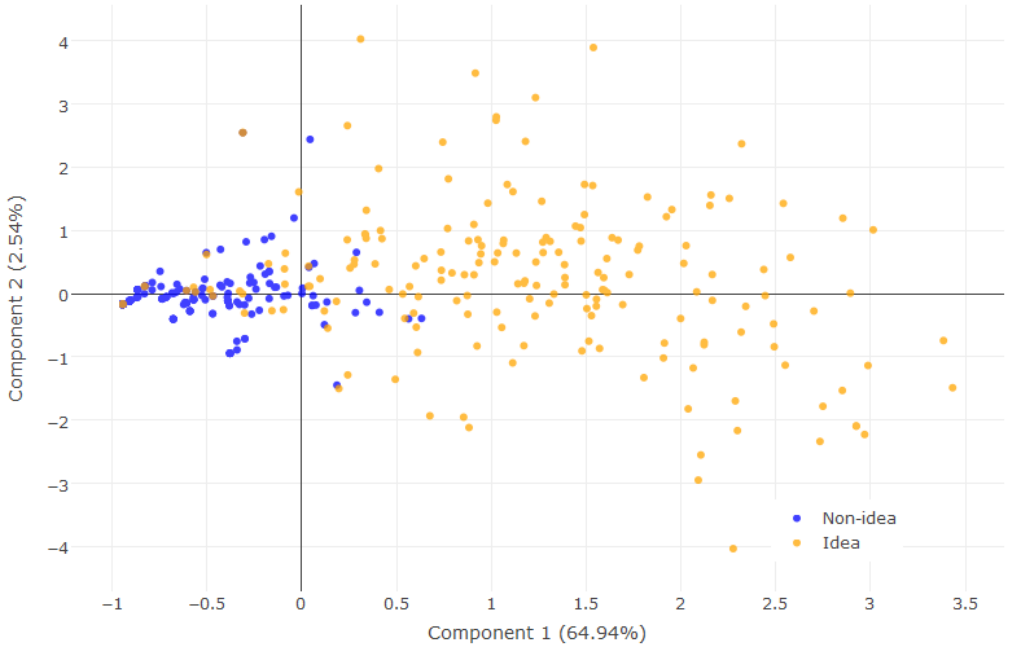


Figure 2 - The figure shows the partial least squares scores based on the beer dataset where short texts and long texts has been removed. The blue dark points are non-idea texts and the orange light squares are idea texts. Component one explains 64.94% of the variance in the target variable. Component two explains 2.54% of the variance in the target variable.

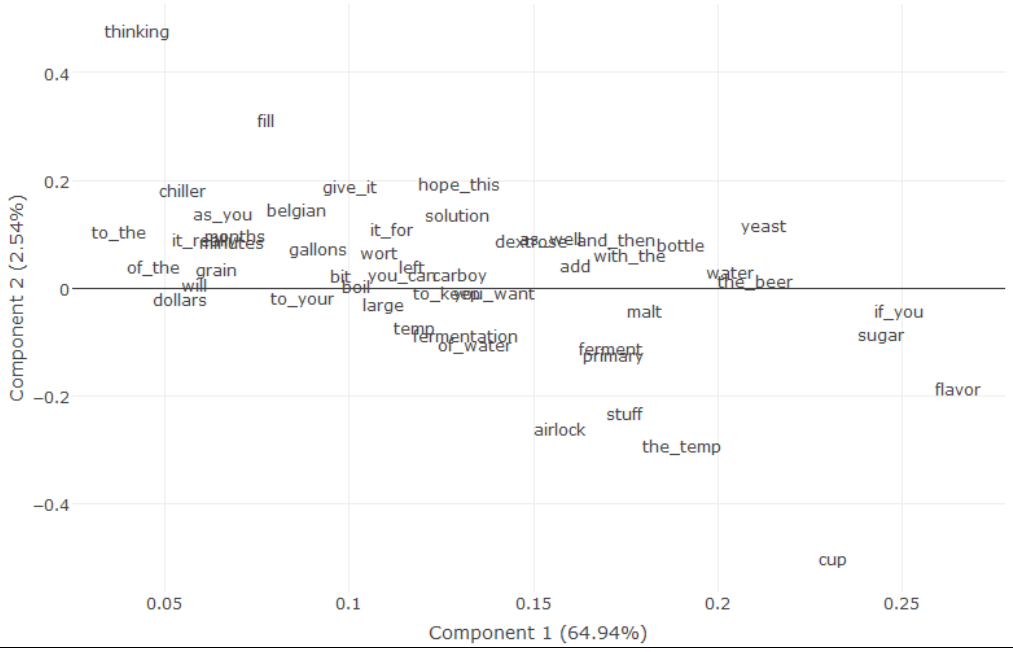


Figure 3 - The figure shows the partial least squares loadings based on the beer dataset where short texts and long texts has been removed. Component one explains 64.94% of the variance in the target variable. Component two explains 2.54% of the variance in the target variable.

Table 9 - Beer idea text containing the term "if you"

´I find that using prime-tabs or carbonation drops or some such works very well...if you add the requisite number just before you cap your bottles, you get a nice uniform carbonation throughout the batch.´

Table 10 - Beer non-idea text

´I have a book on commercial brewing. Unfortunately, I misplaced it when moving house but I'll find it and give more info. However, it also makes common sense. You can't really rack 30,000 litres of beer and prime 80,000 bottles. Even bulk priming is out of the question as it involves racking and stirring in sugar.´

5.2 - Lego community

5.2.1 - Lego classification task

The results from the classification task on the full Lego dataset show that the support vector machine classifier performed slightly better than the partial least squares classifier (Table 11). The support vector machine classifier shows no tendency to over-fit or over-generalize. Neither was this the case for the partial least squares classifier. As with the beer dataset, we do not consider the performance between the two classification techniques remarkably different.

Table 11 - Results of classification task on the full Lego dataset

Technique	Tuning	Partition	Precision	Recall	F_1
Support vector machine	C = 1e-04 Cut-off = 0.43	Validation	0.96	0.97	0.97
		Hold-out	0.96	0.95	0.95
Partial least squares	Components = 2 Cut-off = 0.26	Validation	0.96	0.97	0.97
		Hold-out	0.95	0.92	0.94

When we applied the support vector machine classifier on the Lego doubt set and the Lego hold-out set it was again clear, as expected, that the texts the crowdworkers could not agree on, were equally difficult for the support vector machine to classify (Table 12). Again,

the partial least squares classifier behaved different than the support vector machine classifier and we notice the same two aspects as we noticed in the beer case. The partial least squares classifier, classified 57 out of 78 of the idea texts as doubt texts. Also it classified 1,412 texts as doubt texts whereas the support vector machine classifier classifies only 429 texts as doubt texts (Table 13).

Table 12 - Results from classification task on the Lego doubt set and the hold-out set with the support vector machine classifier

Support vector machine		Classifications		
		Idea (826)	Non-idea (631)	Doubt (427)
Crowd evaluations	Idea (78)	85.9% (67)	3.85% (3)	10.26% (8)
	Almost idea (479)	77.24% (370)	8.35% (40)	14.41% (69)
	Doubtful idea (402)	49.75% (200)	19.15% (77)	31.09% (125)
	Doubtful non-idea (384)	35.16% (135)	33.33% (128)	31.51% (121)
	Almost non-idea (422)	12.09% (51)	64.69% (273)	23.22% (98)
	Non-idea (119)	2.52% (3)	92.44% (110)	5.04% (6)

Table 13 - Results from classification task on the Lego doubt set and the hold-out set with the partial least squares classifier

Partial least squares		Classifications		
		Idea (101)	Non-idea (371)	Doubt (1.412)
Crowd evaluations	Idea (78)	26.92% (21)	0% (0)	73.08% (57)
	Almost idea (479)	11.27% (54)	1.88% (9)	86.85% (416)
	Doubtful idea (402)	4.23% (17)	7.71% (31)	88.06% (354)
	Doubtful non-idea (384)	1.82% (7)	18.23% (70)	79.95% (307)
	Almost non-idea (422)	0.47% (2)	38.63% (163)	60.9% (257)
	Non-idea (119)	0.00% (0)	82.35% (98)	17.65% (21)

When short- and long texts were removed from the Lego dataset, performance of the support vector machine classifier decreased (Table 14). This was also the case for the partial least squares classifier. There is no noteworthy different in the performance between the support vector machine classifier and the partial least squares classifier. None of the classifiers show a tendency to over-fit or over-generalize. The difference in the results obtained on the full set of texts and this reduced set of text show us that the relatively short Lego text and relatively long Lego texts are easier to separate for both of the classifiers than are texts of similar length.

Table 14 - Results of classification task on the Lego dataset when short texts and long texts have been removed from the dataset

Technique	Tuning	Partition	Precision	Recall	F_1
Support vector machine	C = 1e-06 Cut-off = 0.37	Validation	0.75	0.75	0.75
		Hold-out	0.68	0.81	0.74
Partial least squares	Components = 1 Cut-off = 0.20	Validation	0.68	0.81	0.74
		Hold-out	0.63	0.94	0.75

5.2.2 - Lego idea term selection and interpretation

We progressed with the partial least squares classifier for term selection. Here, the significance multivariate correlation term selection procedure resulted in the selection of 50 terms. The scores plots (Figure 4) show a trend that suggests that non-idea texts and idea texts are separated by the two components but mainly by component one, A principal component analysis showed the same trend. We notice that terms like, “would be”, “they would”, “i think”, “idea”, “could be” and “dear Lego” are terms that separate the Lego idea texts from the Lego non-idea texts (Figure 5). The text in Table 15 is an example of a typical idea from the Lego community related to a new product that the community member would like. To us, it is an idea because it reflects suggestion dialogue from one community member to Lego. In Table 16, a non-idea text is shown, and to us this particular text reflects simple information exchange between two community members.

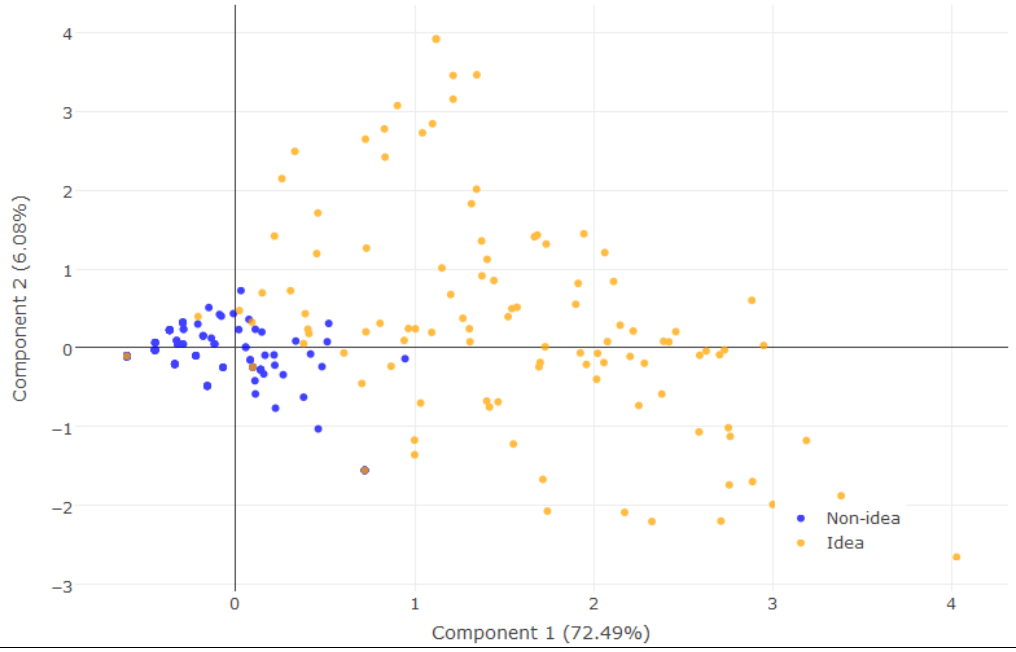


Figure 4 - The figure shows the partial least squares scores based on the Lego dataset where short and long texts has been removed. The blue dark points are non-idea texts and the orange light squares are idea texts. Component one explains 70.20% of the variance in the target variable. Component two explains 5.91% of the variance in the target variable.

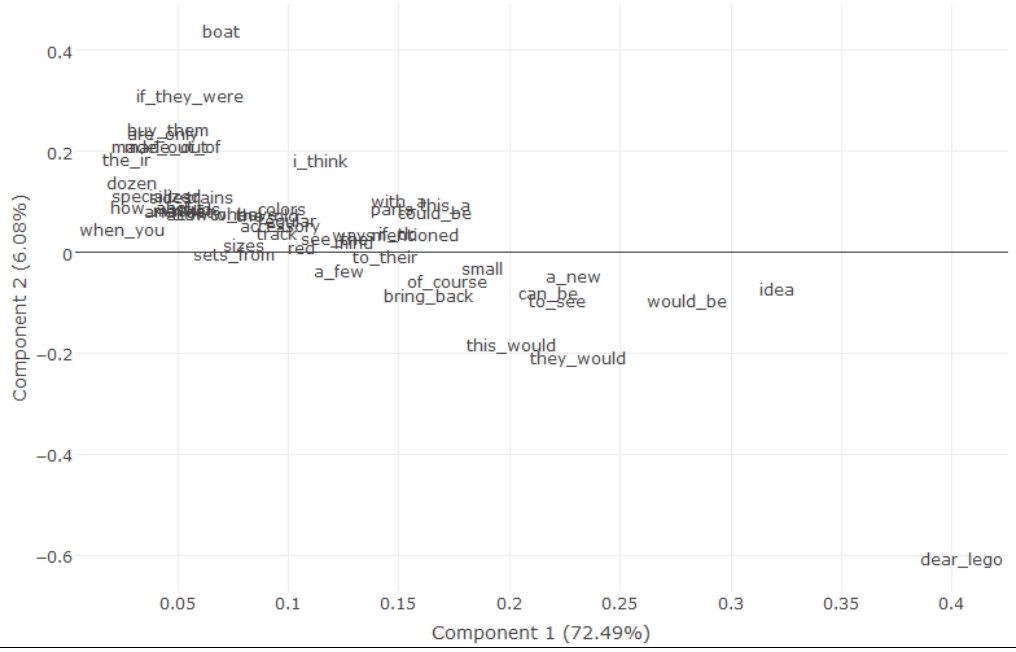


Figure 5 - The figure shows the partial least squares loadings based on the Lego dataset where short and long texts has been removed. Component one explains 72.49% of the variance in the target variable. Component two explains 6.08% of the variance in the target variable.

Table 15 - Lego idea text containing the term "would be"

´Maybe a little over the top or expensive, but a solar powered kit would be interesting. The monorail could run on a smooth strip that clips to the sides of the track like a T-shape track. This could be considered a far more modern and advanced style.´

Table 16 - Lego non-idea text

´Be aware that the event at LLW isn't actually organised by Lego as such. Clearly LLW participate, but the actual event is organised and administered via Red Letter Days. http://www.redletterdays.co.uk/home/index.asp And what you might find even more frustrating is that they apparently cancelled the last one through lack of interest!´

6. Discussion

The main aim of the research presented in this paper was to investigate which terms drive automatics classification of ideas in online communities? What do people tend to say when they write about ideas in online communities? To reach this destination we first investigated if the supervised machine learning technique, partial least squares, can obtain high classification performance on this type of classification task, in comparison to support vector machines. Here, the results of the classification task on each individual idea dataset supports the claim that partial least squares is suited for this type of task. However, the support vector machine classifiers *did* perform slightly better than the partial least squares classifiers. But, because the difference in performance was only minor, it seems like partial least squares can play a role when it comes to interpretation of classification results on this type of text classification task.

By using the partial least squares simultaneous multiple correlation approach for term selection, we were able to find idea predictor terms for each community. We reduced the set of terms from 11,049 terms to 50 terms for the beer dataset, and we reduced the set of terms from 12,426 terms to 50 terms for the Lego dataset. This reduction was done *automatically* as it was purely based on a computational procedure. The identified beer idea terms are terms like “if you”, “solution”, “you want”, “you can”, “thinking”, “the beer”, “sugar”, “flavor” and “yeast”. The identified Lego idea terms are terms like “would be”, “they would”, “i think”, “idea”, “could be” and “dear Lego”. In relation to the “dear Lego” term it is worth to mention that the particular Lego community used for this study is called “Dear Lego” and it is a community specifically meant for new product ideas for Lego. Thus, many community users write “Dear Lego” as the first sentence in a posting before they describe the idea to the new product. Therefore this term is a strong idea predictor for this community.

Upon reflection we see a trend that suggests that idea texts in the brewing community contain two categories of terms: A *suggestion*/*solution* category and a *domain* category. Suggestion/solution terms for the beer community are terms like “if you”, “solution” and “you want”. Domain terms are terms like “the beer”, “sugar”, “flavor” and “yeast”. For the Lego community the idea texts also contain both suggestions/solution terms and domain terms, but it is the suggestion/solution terms that are dominating the classification loadings. This is interesting in relation Poetz & Schreier (2012) who writes that ideas contain *need*- and *solution* information and it gives us reason to believe that the nature of ideas in an online community context, are reflected in suggestion/solution terms like the ones identified. Thus the nature of ideas in the two online communities used for this study, is of a *suggestion/solution* nature. What we mean by this is that idea texts typically reflect dialogue between community members who suggest solutions to problems to each other. And further, in the Lego community, the community members to do not only suggest solutions to problems to each other, *but* to Lego. This observation is interesting because many of the beer ideas are ideas related to how to do something in the brewing process and not directly suggestions for new products. And, many of the Lego ideas on the other hand are actually ideas for new products suggested to Lego. Thus the nature of the ideas in the two online communities are not *only* of a suggestion/solution nature, but the nature of the ideas in the beer community seems to be related to brewing processes’ whereas the nature of the ideas in the Lego community seems to be related to new products that Lego should develop. Thus, the interesting question for future research is if a community like the beer community actually *does* contain any new product ideas?

If future research is to develop similar methods and confirm/disconfirm the results we have reported, we point to one main limitation of our study. This limitation is that in order to ensure that the identified idea texts are in fact idea texts (and the identified non-idea texts are

in fact non-idea texts), we used a very strict exclusion criterion because we only accepted texts where five out of five raters had agreed on the target class. This resulted in the exclusion of many potential training texts. Exactly 1,607 beer texts were excluded and 1,687 Lego texts were excluded. We named these texts *doubt texts* and in comparison Christensen et al. (2016) only excluded 197 texts out of 3,000 training texts. Future research should attempt to generate training datasets that has not been victims to this kind of training data exclusion. One solution to this problem is to frame the classification task as a regression task. In that way, the raters evaluating the texts would have to evaluate each text on a continuous/discrete scale rather than in a binary fashion. Another solution could be to generate the target variable in a 2-step procedure. First, crowdsourcing is used to identify the very clear idea texts and non-idea texts where all raters could agree, as in our study. Secondly, the texts on which the raters could *not* agree are evaluated by trusted colleagues and/or employees.

The results presented in this paper are interesting because they tell us that there is reason to believe that human ideation in an online community context is reflected in specific terms. The practical implications of this discovery is that it may be possible that *one* generic machine learning classifier can used for detecting ideas in *any* online community context. This will be for future research to investigate and for future research we imagine a study, where idea texts and non-idea texts from *many* online communities of different topical nature are used to train, tune and test machine learning classifiers. If such classifiers can be perfected, it opens up for a series of new research questions related to the automatic identification of innovation- and marketing relevant information in online consumer chat. Because if ideas can be automatically identified, is it then also possible to identify consumer *problems,* consumer *needs* and consumer *complaints*? We imagine that these paths will be interesting for future research on predictive text mining in the area of marketing- and innovation management and in food and non-food product domains.

7. Conclusion

The main aim of this paper was to investigate how ideas are expressed in an online community context. What is the nature of ideas in online communities? To reach this destination we first collected ideas from two different online communities. One community related to beer brewing and one community related to the product Lego. We established that partial least squares *can* be used for classification on this particular text classification problem. This is interesting because it supports the claim that partial least squares should be considered for future text classification tasks, as it *also* allows for easy identification of important terms/variables and interpretation of the underlying pattern that is driving automatic classification.

We used partial least squares and the significance multivariate correlation measure to derive the terms (i.e. words and expression) that drive classification of ideas written as text in two online communities. For the beer such terms are terms like “if you”, “solution”, “you want”, “you can”, “thinking”, “the beer”, “sugar”, “flavor” and “yeast”. For the Lego community such terms are terms like “would be”, “they would”, “i think”, “idea” and “could be”. Our results suggest that ideas in the two chosen online communities are reflected in suggestion/solution dialogue. And further, the ideas written in the beer community is often process ideas related to how to do something in the brewing process, whereas the ideas in the Lego community is often new product ideas suggested to Lego.

8. References

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & da Fontoura Costa, L. (2014). A systematic somparison of supervised classifiers. *PLoS ONE*, *9*(4), 1–14.

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

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology*, *609*, 223–239.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory*.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161–168). ACM.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2016). In search of new product ideas: Identifying ideas in online communities by machine and text mining. *Creativity and Innovation Management (In Press)*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

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.

Eickhoff, C., & de Vries, A. P. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, *16*(2), 121–137.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, *3*, 1289–1305.

Franke, N., & Shah, S. (2003). How communities support innovative activities: an exploration of assistance and sharing among end-users. *Research Policy*, *32*(1), 157–178.

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

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning - data mining, inference and prediction* (Second edition). Stanford, CA: Springer.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

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.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 246–268.

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

Lee, G. K., & Cole, R. E. (2003). From a firm-based to a community-based model of knowledge creation: The case of the Linux kernel development. *Organization Science*, *14*(6), 633–649.

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.

Linoff, G., & Berry, M. (2011). *Data mining techniques: For marketing, sales, and customer relationship management* (3. Edition). Indianapolis, IN: Wiley publishing.

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: Exploring Users as a Proxy for Expert Judges. *Journal of Product Innovation Management*, *33*(1), 4–18.

Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. *PLS Methods*, *58*(2), 85–95.

Martens, H., & Næs, T. (1991). *Multivariate calibration*. New York, NY: John Wiley & Sons.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to - Multivariate Calibration and Classification*. Chichester, UK: NIR Publications.

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.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*.

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Radovanović, M., & Ivanović, M. (2008). Text mining: Approaches and applications. *Novi Sad J. Math*, *38*(3), 227–234.

Sautter, G., & Böhm, K. (2013). High-throughput crowdsourcing mechanisms for complex tasks. *Social Network Analysis and Mining*, *3*(4), 873–888.

Ståhle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *Journal of Chemometrics*, *1*(3), 185–196.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Tran, T. N., Afanador, N. L., Buydens, L. M. C., & Blanchet, L. (2014). Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemometrics and Intelligent Laboratory Systems*, *138*, 153–160.

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.

von Hippel, E. (2001). Innovation by user communities: Learning from open-source software. *MIT Sloan Management Review*, *42*(4), 82–86.

Wang, A., Hoang, C. D. V., & Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, *47*(1), 9–31.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2. edition). San Francisco, CA: Morgan Kaufmann publishers.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In B. Kågström & A. Ruhe (Eds.), *Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22–24, 1982* (pp. 286–293). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130.

Zanasi, A. (2007). *Text mining and its applications to intelligence, CRM and knowledge management* (1. edition). Southampton, UK: WIT Press.

Paper 4

Kasper Christensen, Joachim Scholderer, Stine Alm Hersleth, Tormod Næs, Knut Kvaal, Torulf Mollestad, Nina Veflen and Einar Risvik

How good are ideas identified by an automatic idea detection system? (Submitted for publication in Creativity and Innovation Management)

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

Creativity & Innovation Management: Special issue on Big Data for Open Innovation

Kasper Christensen[1,2,*], Joachim Scholderer[3,4], Stine Alm Hersleth[2], Tormod Næs[2,5], Knut Kvaal[1], Torulf Mollestad[6], Nina Veflen[2,7], Einar Risvik[2]

[1] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway
[2] Nofima A/S, Ås, Norway
[3] Department of Economics and Business Economics, Aarhus University, Denmark
[4] CCRS and Department of Informatics, University of Zurich, Switzerland
[5] Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Denmark
[6] Altran, Norway
[7] BI Norwegian Business School, Norway
* Corresponding author. Email: kasper.christensen@nofima.no, Tel.: (+47) 94 15 89 93

Word count: 6469

Acknowledgements

The authors would like to thank *Nøgne Ø* and Tom Young, Ingebjørg Christina Nybø and Andrew Windtwood for their help in the data collection. Also, the authors would like to thank *The Foundation for Research and Levy on Agricultural Products in Norway* for funding this project.

Abstract

Online communities are an attractive source of potential ideas for products and process'. Recent advances in machine learning have made it possible to screen the vast amounts of information in online communities and automatically detect user-contributed ideas. However, it is still uncertain whether the ideas identified by such a system 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 (2016; 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 innovation 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

The digitalisation of business life is progressing: more and more tasks can be solved by automated systems. Whilst in the past, these were predominantly tasks of a mundane and repetitive nature, recent advances in artificial intelligence have also made it possible to solve complex problems. A common problem during the introduction of such systems is that they can be intransparent to their prospective users. Whilst the traditional business processes they are intended to rationalise have often been in use for many years and are implicitly trusted by management and staff, newly introduced automated systems lack such a track record. Scepticism and reactance can be the consequence.

To earn the trust of prospective users, automated systems have to enable superior performance. Benchmarked against the traditional business processes they are intended to rationalise, they should lead to increases in effectiveness or efficiency. This is easily demonstrated in application areas such as sales forecasting or inventory control where commonly accepted and routinely measured performance criteria exist. Such criteria rarely exist in more complex and creative areas such as innovation management. The aim of the research presented here is to show how the performance of automated systems in such areas can be evaluated. We will demonstrate this in the context of a particular type of task: the automated detection of ideas for product and process innovations in the contributions to an online developer forum.

Online communities as idea reservoirs

Firms need a continuous stream of ideas to fuel their innovation processes (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, & PJ de Jong, 2011; Poetz & Schreier, 2012; Majchrzak & Malhotra, 2013; Magnusson, Wästlund, & Netz, 2014). Online communities where users exchange experiences and discuss potential improvements are a particularly rich reservoir of ideas for product and process innovations.

Prominent examples include the user 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 exists 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” that have been developed in a distributed manner, utilising online collaboration platforms 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 natural language processing and machine learning methods, Christensen, Nørskov, Frederiksen, & Scholderer (2016) 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 et al. (Submitted manuscript) 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. Since the presence of such linguistic markers can easily be detected in a given post or comment, it can also be exploited for the screening of arbitrarily large collections of posts, comments or other types of semi- or unstructured text. Implemented as a screening tool in a firm'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.

How good are automatically detected ideas?

A crucial question is whether the ideas detected by such an automated system would also be seen as sufficiently novel, feasible and valuable by the R&D or marketing staff who would have to decide if the ideas should be taken further (e.g., developed into concepts or prototypes). Ideas identified by the Christensen et al. (2016) method, for example, have not yet been evaluated by company-internal R&D or marketing staff. The aim of the present paper is to fill this gap. 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. (2016), 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.

These issues reflect potential acceptance problems that were in the innovation literature initially seen as general 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 (e.g., see (Kristensson et al., 2004; Magnusson, 2009; Poetz & Schreier, 2012; Magnusson et al., 2014) and therefore deserve to be given a fair chance. As a consequence, dedicated crowdsourcing systems have gained widespread acceptance in the business community. Our study extends this question to the *mode* of idea harvesting: can user-contributed ideas identified by an artificial intelligence system reach sufficient recognition among company professionals? An online community

related to craft brewing was used as the idea base for our study. Employees of Norwegian craft brewery *Nøgne Ø* evaluated the automatically extracted ideas.

Method

Machine learning for idea detection

The machine learning system we employed is described in detail in Christensen et al. (2016) and Christensen et al. (Submitted manuscript). 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. We expected ideas for product and process to be available in this community. At the time the texts were extracted, the community contained altogether 10582 posts. 3000 of these were extracted for the development of the training of the system. Those that contained ideas were identified by 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. In this process, 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 process resulted in a dataset consisting of 10514 terms representing 10582 texts.

The 3000 training texts were separated from the remaining 10582 texts. From the 3000 training texts, we excluded all texts where not all five CrowdFlower raters had agreed on the

class membership. After excluding these, the new training set contained 1393 texts. 405 of the texts were idea texts and 988 were non-idea texts. The training 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 the tuning of the machine learning system (in the validation set) and for an unbiased evaluation of its performance in the context of previously unseen data (hold-out set). Based on the training set, validation set and hold-out, the automatic idea detection system was trained and tested. The system was based on a linear support vector machine classifier (SVM; for details, see Christensen et al., 2016). Performance statistics are reported in Table 1.

Table 1 - Performance of the automatic idea detection system used by Christensen et al. (Submitted Manuscript)

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

From the remaining 7582 texts which had not been involved in the training, validation and testing of the system in the study by Christensen et al. (Submitted manuscript), another 200 were extracted for the present study. Using the SVM classifier, the texts were scored as to how likely they were to contain an idea. A histogram of the resulting posterior probabilities is shown in Figure 1. These 200 texts were then used in the present study as the idea and non-idea texts to be classified and rated by two brewing professionals.

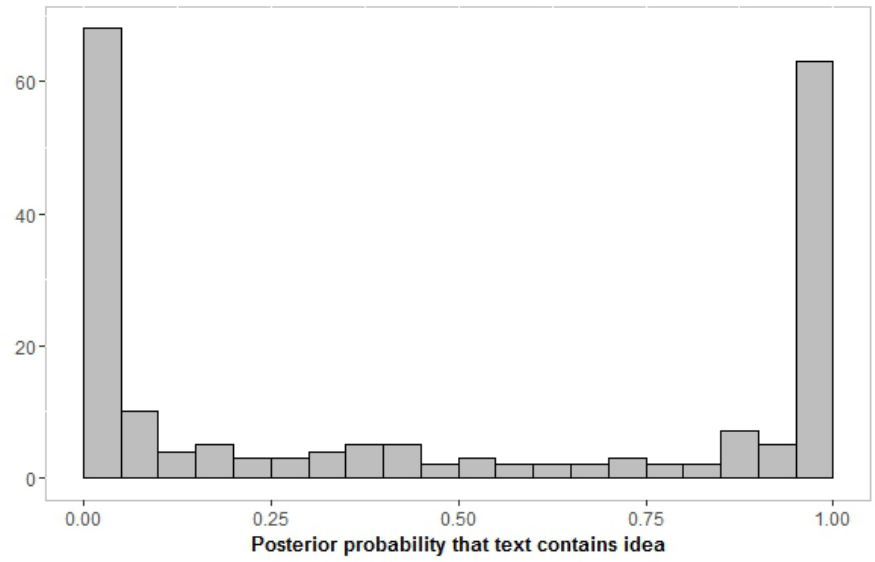


Figure 1 - Histogram of the posterior probability scores generated by the SVM-based automatic idea detection system for the 200 texts used in the present study

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 could 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, & 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 & 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 had 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 valuate 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 (Cohen, 1960; Landis & Koch, 1977; von Eye & von Eye, 2008). 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.

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!´

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 $F_1 = 0.75$. 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 - 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

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.

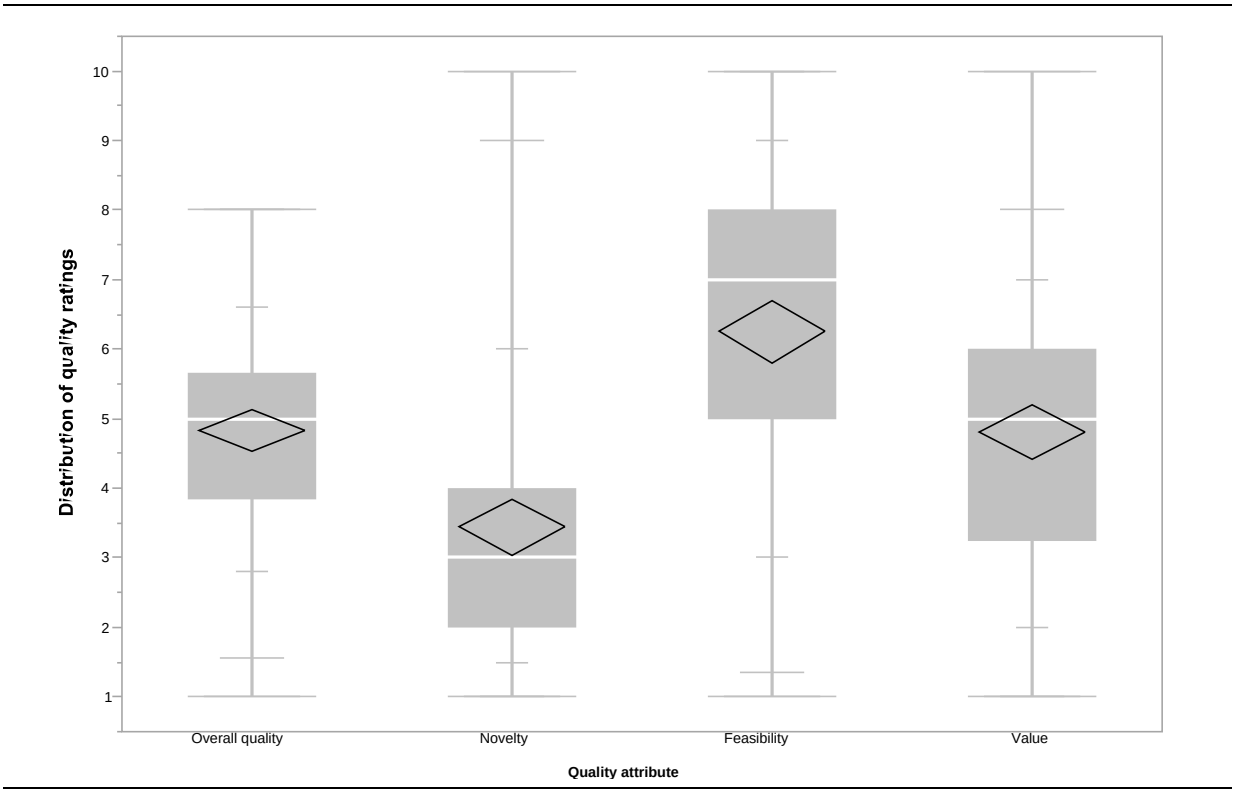


Figure 2 - Box plots of the distribution of quality ratings (overall quality = unweighted average of novelty, feasibility and value; diamonds represent 95% confidence intervals around distribution means)

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.

Discussion and conclusion

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 et al., 2016) would also be regarded as ideas by company-internal staff who will 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.

Since the two experts also differed in terms of their functional responsibilities in the company, it might not even be appropriate to look for perfect agreement—after all, a company's ability to integrate different functional perspectives is one of the strongest predictors of innovation success (e.g., see Evanschitzky, Eisend, Calantone, & Jiang, 2012). 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. (2016) 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 is 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 value 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 et al. (2016) 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.

An interesting detail 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) and Magnusson (2009) for user ideation in general. 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 - 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....´

The results presented here are an evaluation of a particular automatic idea detection system (the one developed by Christensen et al., 2016) 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 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.

References

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.

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. (Submitted manuscript). Mining online community data: The nature of ideas in online communities. *Food Quality and Preference*.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2016). In search of new product ideas: Identifying ideas in online communities by machine and text mining. *Creativity and Innovation Management (Available Online)*.

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.

Evanschitzky, H., Eisend, M., Calantone, R. J., & Jiang, Y. (2012). Success factors of product innovation: An updated meta-analysis. *Journal of Product Innovation Management*, *29*(S1), 21–37.

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.

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.

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: Exploring Users as a Proxy for Expert Judges. *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.

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*.

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.