Contents lists available at ScienceDirect

Food Quality and Preference

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

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

Kasper Christensen[a,b,*], Kristian Hovde Liland[b,c], Knut Kvaal[a], Einar Risvik[b], Alessandra Biancolillo[b,d], Joachim Scholderer[e,f,g], Sladjana Nørskov[h], Tormod Næs[b,d]

[a] *Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway*
[b] *Nofima A/S, Ås, Norway*
[c] *Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway*
[d] *Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Denmark*
[e] *School of Economics and Business, Norwegian University of Life Sciences, Ås, Norway*
[f] *CCRS and Department of Informatics, University of Zurich, Switzerland*
[g] *Department of Economics and Business Economics, Aarhus University, Denmark*
[h] *Department of Management, Aarhus University, Denmark*

ARTICLE INFO

Keywords:
Machine learning
Text mining
Natural language processing
Online communities, ideas
Partial least squares
Support vector machines

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 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. In addition 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

1.1. Background

Online communities can be important drivers of knowledge generation for the firm. They allow people with similar interests to gather and interact. Thus, online communities become locus points for people all over the world that can 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 (Jeppesen & Frederiksen, 2006; Lee & Cole, 2003; von Hippel, 2001). Facebook groups, google forums and newsgroups are all examples of online community types.

A special kind of knowledge that has occupied innovation management scholars and R & D people is *ideas* (Dean, Hender, Rodgers, & Santanen, 2006; Kristensson, Gustafsson, & Archer, 2004; Magnusson, 2009; Magnusson, Wästlund, & Netz, 2014; van den Ende, Frederiksen, & Prencipe, 2015). Ideas represent a 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). To secure a continuous stream of ideas some firms have established their own online communities, where dedicated product users and consumers gather to discuss- and suggest ideas to the firm (e.g. Dell (di Gangi, Wasko, & Hooker, 2010) or Propellerhead (Jeppesen & Frederiksen, 2006).

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. Another widespread type of community exists, namely the type that is established by the users of the community itself. This type of community

[*] Corresponding author at: Nofima A/S, Ås, Norway.
E-mail address: kasper.christensen@nofima.no (K. Christensen).

http://dx.doi.org/10.1016/j.foodqual.2017.06.001
Received 2 December 2016; Received in revised form 24 May 2017; Accepted 5 June 2017

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 (Antorini, Muñiz, Albert, & Askildsen, 2012; Franke & Shah, 2003; Füller, Jawecki, & Mühlbacher, 2007).

As opposed to the firm-hosted communities related to 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, 2017). 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 text classification problems. The downside of using support vector machines is that they are non-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 a problem when we, the users of the method, seek to explain the underlying phenomenon that enables automatic classification. And, if future research want to aim at improving data representation and methodology on text classification problem, it is important that future methods are designed in a way that gives insights into relations that drives classification.

1.2. Aims and scientific contributions

The present paper has two aims: The first and primary aim is an investigation of whether a well-known method in the area of sensory- and consumer science, *Partial Least Squares* (see e.g. Martens & Næs, 1991; Wold, Martens, & Wold, 1983) can provide the additional interpretation power that the support vector machine lacks. Partial least squares regression is a method that has proven to be useful for classification as well as for interpretation of the relations that drives classification. It has however, not yet been applied for automatic idea identification in online communities. We see a room for investigating whether that 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 secondary aim is to extend the approach used in Christensen et al. (2017) 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 machine learning classifiers were in doubt. Two very different online communities cases, Lego and beer brewing, will be used for evaluating the methodology.

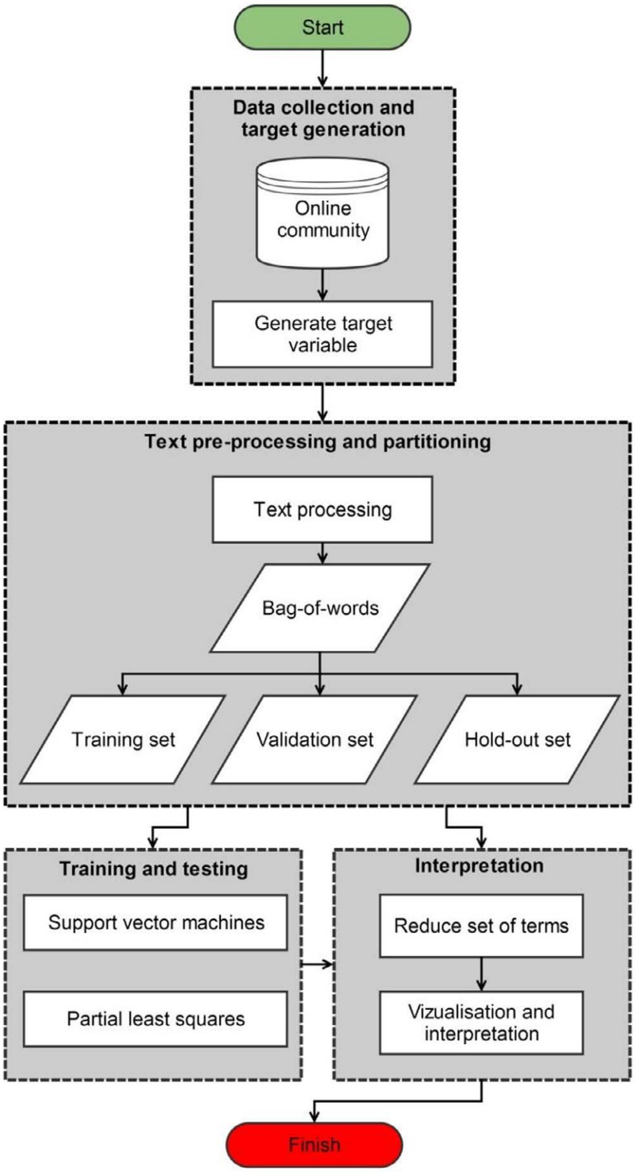

Fig. 1. Flowchart showing the supervised machine learning procedure for classification and interpretation that we apply.

2. Choice of methods

Our supervised machine learning procedure for idea detection can be divided into four main parts (See Fig. 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. The target variable contains the information the machine learning technique uses for learning. To generate a target variable for text classification tasks, crowdsourcing can be used (Christensen et al., 2017; Howe, 2006; Wang, Hoang, & Kan, 2013). When utilizing crowdsourcing for this type of task, Sautter and 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.,