creativity and innovation management

In search of new product ideas: Identifying ideas in online communities by machine learning and text mining

Journal:	*Creativity and Innovation Management*
Manuscript ID	1272-CIM-A-10-2015.R2
Manuscript Type:	Article
Keywords:	Idea mining, Machine learning, Text mining, Natural language processing, Online communities, Innovation management, Marketing management

SCHOLARONE™
Manuscripts

For Peer Review

In search of new product ideas: Identifying ideas in online communities by machine learning and text mining

Abstract

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 analyzing 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.

For Peer Review

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; also see 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., von Krogh, Spaeth, & Lakhani, 2003; Hertel, Niedner, & Herrmann, 2003; Henkel, 2006; Foss, Frederiksen, & Rullani, Forthcoming). Similarly, in the marketing literature, online communities built around particular brands, e.g. Audi or Lego (Füller, Bartl, Ernst, & Mühlbacher, 2006; Antorini, Muñiz, & Askildsen, 2012), are considered important sources of ideas for brand or line extensions and product re-positioning (Ogawa & Piller, 2006).

For Peer Review

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 text and requires substantial pre-processing before it can be statistically analyzed (Netzer, Feldman, Goldenberg, & Fresko, 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 since 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 including ideas from texts not including 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 & Salge (2015) described 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

For Peer Review

allows for automatic mapping of the topic structure in a collection of texts. Tirunillai & Tellis (2014) used the same technique to describe the topic structure in a collection of online product reviews. Also, Kaplan & Vakili (2014) adapted the latent Dirichlet allocation for the purpose of measuring degrees of novelty in collections of patents. Thorleuchter, den Poel, & 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 multi-dimensional 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.

For Peer Review

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, Hender, Rodgers, & Santanen, 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, & Nunamaker (2007) compare different ways of scoring such data to evaluate the effectiveness of idea generation techniques. They recommend 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 if 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 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.

For Peer Review

--- Table 1 ---

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 aim at *discovering* patterns in data and represent them in a low-dimensional form, often accompanied by visualisations 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 do not allow for making *distinct* binary predictions (i.e. idea vs. non-idea) and are therefore inherently *descriptive*. This is important, since all existing applications of text mining that we reviewed above (Antons et al., 2015; Kaplan & Vakili, 2014; Netzer et al., 2012; Thorleuchter et al., 2010) 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).

For Peer Review

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 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 nonlinear, 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.

For Peer Review

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 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, 2015), and they have generated new products and new products lines and created new market opportunities for Lego (Antorini, Muniz & 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.

For Peer Review

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 $\kappa(\text{max}) = 0.87$. Hence, the observed value of kappa was 63% of its maximum value, moving it into a range that can be regarded as substantial.

For Peer Review

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 whitespaces 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. (2015) used a sparsity threshold of 0.1%. In the present analysis, we experimented with sparsity thresholds of 2.5%, 1%, and 0.25%.

For Peer Review

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 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, naive 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.

For Peer Review

--- Table 2 ---

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, Fernandez, Barrenechea, Bustince, & Herrera, 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.

For Peer Review

Assessment of classification performance

The performance of the competing classification models will be compared using measures derived from signal detection theory (see Witten & Frank, 2005; Stanislaw & Todorov, 1999). 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 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:

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

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

$$F_1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall}+\text{Precision}} \tag{3}$$

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:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

For Peer Review

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

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 et al., 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 had 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.

For Peer Review

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 Figure 1 (grouped by input data and hyper-parameters 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

For Peer Review

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.

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.

--- Figure 1 ---

--- Table 3 ---

--- Table 4 ---

For Peer Review

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 syntax level, following a specific pattern, when it leaves the human mind to be written in 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.

For Peer Review

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 et al. (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 & 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 et al., 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 *if* 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 also relative 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 6,500 USD. They assessed 3,000 texts each and they identified 137 ideas. This corresponds to a price of 47.45 USD per idea and 2.17

For Peer Review

USD per text assessment. The costs of identifying 100 ideas would then sum to 4,745 USD if no pre-filtering were applied. If, on the other hand, our method were applied as a pre-filter, the identification of 100 ideas would cost 238 USD 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 favor of the F_1-measure. However, one can imagine cases where a researcher or a firm would favor 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 favors 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 favoring 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 Peer Review

--- Table 5 ---

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.

For Peer Review

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, 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 6,500 USD 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 requires 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 classifier performed mediocre on precision, recall, and F_1-measure 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.

For Peer Review

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 neighbors, 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 understandable by people with little knowledge of machine learning and text mining. We *could have* chosen to show the performance of these classifiers also, 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 disconfirming 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 medias, 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 catalysts for industry and firm innovation?

For Peer Review

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 if 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 sign of over-generalization or 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.

For Peer Review

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. (2015). 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*, In press.

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.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine Learning Algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1).

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371–1391.

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

Besemer, S., & O'Quin, K. (1999). Confirming the Three-Factor Creative Product Analysis Matrix Model in an American Sample. *Creativity Research Journal*, *12*(4), 287–296.

Besemer, S. P. (1998). Creative product matrix analysis: Testing the model and a comparison among products - Three novel chairs. *Creativity Research Journal*, *11*(4), 333–346.

For Peer Review

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine Learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), 61–70.

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.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

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

Dahlander, L., & Piezunka, H. (2015). Distant search, narrow attention: How crowding alters organizations filtering of suggestions in crowdsourcing. *Academy of Management Journal*, *58*(3), 856–880.

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.

Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, *10*(5), 1048–1054.

Foss, N. J., Frederiksen, L., & Rullani, F. (Forthcoming). Problem-formulation and problem-solving in self-organized communities: How modes of communication shape project

For Peer Review

behaviors in the free open-source software community. *Strategic Management Journal*.

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.

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.

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, *35*(7), 953–969.

Hertel, G., Niedner, S., & Herrmann, S. (2003). Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research Policy*, *32*(7), 1159–1177.

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.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137–142). Presented at the C. Nédellec & C. Rouveirol (Eds.), Proceedings of 10th European Conference on Machine Learning (ECML-98), Chemnitz, Germany.

For Peer Review

Kaplan, S., & Vakili, K. (2014). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 1436–1457.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

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

Kudrowitz, B. M., & Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design*, *24*(2), 120–139.

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), 159–174.

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.

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

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.

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*.

For Peer Review

Ogawa, S., & Piller, F. T. (2006). Reducing the Risks of New Product Development. *MIT Sloan Management Review*, *47*(2), 65–71.

O'Quin, K., & Besemer, S. P. (2006). Using the Creative Product Semantic Scale as a Metric for Results-Oriented Business. *Creativity and Innovation Management*, *15*(1), 34–44.

Pang, B., Lee, L., & 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* (Vol. 10, pp. 79–86). Association for Computational Linguistics.

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.

Reinig, B., Briggs, R., & Nunamaker, J. (2007). On the Measurement of Ideation Quality. *Journal of Management Information Systems*, *23*(4), 143–161.

Sawhney, M., & Prandelli, E. (2000). Communities of creation: Managing distributed innovation turbulent markets. *California Management Review*, *42*(4), 24–54.

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.

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.

For Peer Review

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

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.

Wallas, G. (1926). *The art of thought*. New York, USA: Solis Press.

Wasko, M. M., & 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., & Provost, F. (2001). *The effect of class distribution on classifier learning: an empirical study* (Technical report No. ML-TR-44). Newark, NY: Department of computer science, Rutgers university.

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

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535.

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

Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, *3*(4), 243–269.

For Peer Review

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 web site 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?? -- ##COMAPNY 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!"

For Peer Review

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 neighbor (3) Decision trees (4) Neural networks
Bijalwan, Kumar, Kumari & Pascual (2014)	Comparative study	Text data	(1) Nearest neighbours (2) Naïve Bayes (3) Alternative technique
Baharudin, Lee, & Khan (2010)	Review of supervised learning techniques for text mining		(1) Support vector machines (2) Nearest neighbor (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 neighbor
Kotsiantis, Zaharakis & Pintelas (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 neighbor
Pang, Lee & Vaithyanathan (2002)	Comparative Study	Text data	(1) Support vector machines (2) Alternative method (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 neighbor (3) Alternative technique (4) Decision trees (5) Naïve Bayes

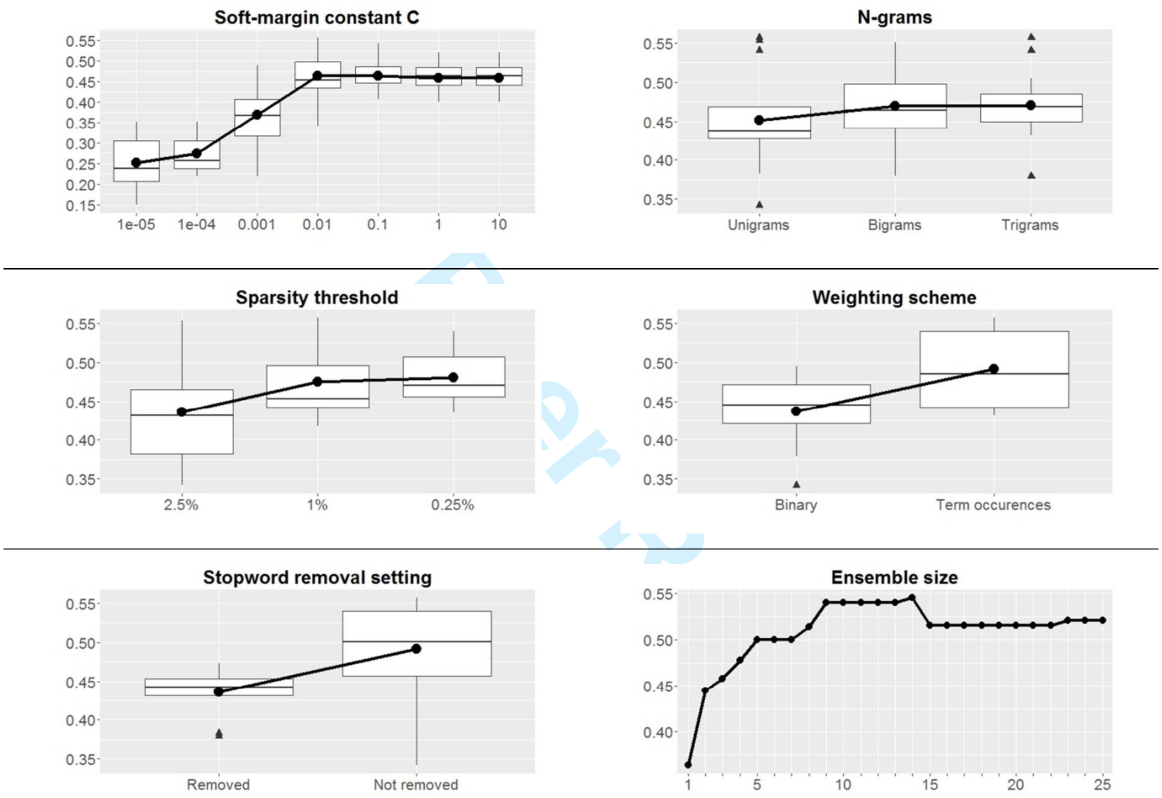

Figure 1 – Distribution of classification performance (F_1-measure) in the validation set as a function of soft-margin constant C, order of n-grams included in the representation, sparsity threshold, weighting scheme, stopword removal setting, and ensemble size.

Table 3 – Classifier tuning with respect to cost (C)

C	*TP*	*TN*	*FP*	*FN*	*F*	*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

Table 4 – Classifier performance on validation set and external (Crowd) test set

Set	*TP*	*TN*	*FP*	*FN*	*F*	*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

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##"

For Peer Review

In search of new product ideas: Identifying ideas in online communities by machine learning and text mining

Dear Editors,

First, thank you for the opportunity to resubmit our paper to CIM. We appreciate all the comments and suggestions for changes we received. These has served as a *huge* source of support for improving, not only this paper but also influencing other parts of our work portfolios.

With the current version, we feel the paper offers a clear account of an important piece of research.

Below you find all of the reviewer comments and questions *and* our answers.

Reviewers comment

As the reviewer very correctly points out you have substantially improved your paper. However, as the method is quite new you have to convince the reader and the scientific community of the validity of your method. Therefore, I would like to ask you to provide quite some more information about the testing of the classifiers (method and results).

Answer to comment

We realize that we made a lot "choices" when we decided how to model the text data and we realize that the results we report do not justify these choices sufficiently. The argument for this choice were, as you might recall, that this paper was never intended to be a 'traditional' machine learning paper. *But* rather it was meant as a paper that shows a way of thinking that can be used for creativity and innovation management research. Namely how scholars and practitioners can generate target variables for machine learning that captures interesting patterns such as ideas. *However*, we do recognize that it is always useful to report and share results for future research and we *do* appreciate that the review team have helped us finding the "golden line" for how much to report and not report

Therefore, to meet this comment we added a figure (Figure 1) where we show increase/decrease in classifier F-measure performance by changing the levels of C, stopword removal, sparsity levels, n-grams and weighting scheme and ensemble size. Hopefully this makes the reasons for our modelling choices more transparent and can also support future research in developing similar methods. Maybe even over time, a *golden practice* for supervised machine learning for text mining, in our research field, can be developed.

Reviewers comment

Please elaborate a lot more on the precision of your results, not only the LEGO community but also the crowdsourcing community (which one?). While I do acknowledge that you are using a new method you have therefore specifically show the reader that your results are trustworthy.

Answer to comment

We see two questions in this comment.

For Peer Review

The first question is related to what community we used for the external validation study. We understand that there is some confusion about whether or not the data the in *external test set* was from the same community, because the *idea rate* in this new external set, seems to be different from the idea rate in the training set and validation set. It *is* the same community. Therefore, to meet this comment we have specified this in the text.

The second question is related to the precision of the results and we do agree that it is appropriate and interesting to discuss the precision of the results more, as well as relating it to other studies. Therefore we have elaborated on the precision in the discussion as well as we have added Cohens Kappa as a performance measure. As you will see, adding Cohens Kappa allows us to compare our results directly with the Von Eye benchmark scale we refer to in the paper. Further, we compare with previous similar studies. We argue that our results are satisfactory.

Reviewers comment

Although you point out in your response to the reviewers that your text addresses are very specific target group and therefore proof readers are hard to find, I do not agree. There are quite some substantial grammatical and spelling errors in your text and I expect you to have your paper proofread by a native speaker before your resubmit it.

Answer to comment

As a response to this comment, we have had a professional native English-speaking proofreader to go through our paper before resubmitting it. We feel the paper is crisp and flows well in the current version. And, we hope that this effort has resolved this issue.

Reviewers comment

Concerning stop word removal, you wrote: "stopword removal had a negative effect on performance. Therefore, we excluded it as a parameter from the tuning procedure reported in the manuscript." What do you mean by a negative effect on performance? Which performance, how was it measured? What was the effect like? Please show me!

Answer to comment

By negative performance we mean that *on average* the classifiers F1-measure performance decreases when stopwords are removed. This we show in Figure 1. We also highlight in the "Training and tuning procedure" section that it is the F1-measure we use as performance metric when modelling and tuning the classifier.

Reviewers comment

Concerning n-grams:
You state that you counted n-grams. But my question was how did you come up with the list of n-grams that you counted. Did you simply try out all possible combinations of n-grams?

Answer to comment

For Peer Review

Yes, we generated all possible combinations of n-grams and *no* list was used. The n-grams that were left in the analysis were the ones that would fall inside the defined sparsity threshold. To meet this comment we have specified in the text, that we identified all possible combinations of n-grams. Also in figure 1 we show the F-measure effects of changing the n-grams levels.

Reviewers comment

Concerning your answer on sparse terms:
Again, how did you measure the performance of the classifier / the effect of sparse word removal on the performance of the classifier?
I understand that you did not include your tests into the paper, but to convince the review team of your approach, it would be helpful to share the results.

Answer to comment

This is the same answer as with stopwords. By negative performance we mean that *on average* the classifiers F1-measure performance increased when the sparsity level is increased. We show this in Figure 1. We also specify in the "Training and tuning procedure" section that it is the F1-measure we use as performance metric when modeling and tuning the classifier.

Reviewers comment

Concerning the choice of your classifier:
I agree that your paper has a focus on showing that classifiers can be used in innovation research. However, I still feel that you should at least add a table that introduces the methodological class of classifiers more broadly showing different classifiers, their scientific sources, their methodological differences etc. to show why your chosen classifier is the right one in order to apply it for automatic idea extraction. I mean, in the end your aim is to convince other scholars and practitioners that your method is the right one and that they should follow you and cite you. To enhance their trust in your method, you should inform and educate them more and show that you indeed made the right decision.

Answer to comment

We agree that the arguments for choice of classification algorithm is relevant and we appreciate that this has been pointed out. To meet this comment we have conducted a literature review aimed at comparing well-known supervised machine learning techniques. The results of this review results we put into a new table (Table 2). We hope that it is now clear that for text mining tasks, the support vector machine is a *very* reasonable choice when the aim is to achieve high predictive performance.

Reviewers comment

Concerning splitting the dataset:
I think it is a bit misleading that while you now state in the text on signal detection theory that splitting into three sets is necessary, you split only into two sets. I think you should explain you logic a bit more by referring to the external set as the third one (as you did in the response).

Answer to comment

For Peer Review

Yes, indeed this need to be more clear. Thank you for pointing that out. We have added a sentence specifying this in the paragraph.

Reviewers comment

Concerning the external validity of your approach:
I think your approach is good using new texts being classified and using crowdsourcing (I imagine Mechanical Turk?) to classify them. However, I wonder why the ratio of ideas is so much higher in your new texts (50%) than in your original dataset (4.9%). Has the community changed?
Moreover, I think you should put your results into relation to other classifier studies. Are your results good (or not)? I mean, having a precision of only 0.41 seems rather low, meaning that 59% percent of all ideas being classified as ideas are in fact no ideas. While this might be ok from a practitioner's view (one has to go through the classified texts again and see whether they contain ideas), missing true ideas is a real problem. What if a groundbreaking breakthrough idea is missed? Here, a recall of 79% seems also quite low. If I use your method, I will miss 21 out of 100 ideas, perhaps missing all the groundbreaking ideas leaving only the minor ideas. Looking at the validation study with new ideas, the results get even worse. Hence, tuning the method with regards to the F measure seems wrong to me in this specific application. What do you think?

Answer to comment

We see three separate questions in this comment. Therefore we have separated the answer into three parts.

The first question is related to the amount of ideas in the training/validation set vs. the amount of ideas in the external set. Our answer to this comment is:

We understand the confusion but *no*, we did not change the community. It is still the same community. Notice that we used the classifier to extract 250 idea texts and 250 non-idea texts. This does not suggest anything about the *true* "idea rate" of the community. To meet this comment we have elaborated on this in method.

The second question is related to the relative *goodness* of the results. Our answer to this comment is:

We agree that it would be very useful and interesting to compare the results in our study, with results achieved in other studies of similar nature. Therefore, to meet this comment we included Cohens Kappa as a performance measure. This enables us to benchmark up against the Von Eye benchmark scale (As referenced in the manuscript). Also, we compare with studies of similar nature (2 x Thorleuchter et al) and we compare with the text mining studies identified in our review.

We hope that it is now visible that our results *are* satisfactory.

The third question is related to the use of recall, precision and F1-measure as performance measures. Our answer to this comment is:

Yes, there is definitely a valid discussion in whether or not the F1-measure is the best performance measure. This is however very subjective and our argument for focusing on the F1-measure is that it balances precision and recall, and is the most "neutral" measure and thus also the most appropriate measure to report when doing research.

For Peer Review

We do however acknowledge that recall might in some cases be a better performance measure. In our study we can tune the classifier so that recall gets as high as 94% meaning that one would only miss 6 out of 100 ideas. However, this means that precision drops to 0.20. As an answer to this comment we included a paragraph about this in the discussion *but* we stick to the F1-measure as our main performance measure for tuning because it is the most neutral.

Reviewers comment

Moreover, I think you should explain to reader not having a technical background, that you tweaked your parameters based on your manual classification. As you show, the performance of your technique depends on the choice of C. You need to say that readers cannot take your algorithm and the optimal C of 0.01 to classify their texts since classifiers relay on initial manual training, which is highly case-specific (here: ideas from the Lego community). You should explain that the C found by you only seems to be "true" for the Lego community (and perhaps very similar texts in other toy related idea communities). Here, you need to educate the reader more on the usefulness of classifiers. Only if the text volume is expected to grow (or you have a small training set that is used to learn about many more texts), training a classifier is of worth – at least as far as I see it. Here, you need to discuss what the implication for other scholars or practitioners is. How can they come up with the best value for C given that they have not coded their dataset manually to test it. You need to educate them that they have to do it far a fair share of their texts to train the classifier according to their data.

Answer to comment

Yes, this is *indeed* important to point out. In order to meet this comment we write in the support vector machine part that C is ***data dependent*** and need be found computationally for all new data sets.

Reviewers comment

Regarding the paper:
I think your description (or better distinction) of supervised and unsupervised approaches can be a bit more detailed. Have a look at the following article: Bao Y, Datta A. 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science 60(6): 1371–1391. Moreover, it is not true that unsupervised algorithms are not able to classify new data – there are many algorithm that learn online, have a look at online topic modeling, for instance, which is capable of modeling texts that are streamed so that the underlying corpus grows continuously.

Answer to comment

Thank you for yet another very interesting and relevant source. We have incorporated the source in the paper.

Further, we understand that latent Dirichlet allocation can be used on *new* data, in the same way as can cluster analysis and principal components. Therefore we have removed this formulation from the paper as well as we have emphasized that the aim of using supervised machine learning is to *distinctly* classify texts into categories based on a target variable.

For Peer Review