Can AI Help in Crowdsourcing?
Testing Alternate Algorithms for Idea Screening in Crowdsourcing Contests

Jason Bell
Saïd Business School
University of Oxford
Park End St, Oxford OX1 1HP, United Kingdom
Jason.Bell@sbs.ox.ac.uk

Christian Pescher
Friedrich-Alexander-University Erlangen-Nuremberg
Lange Gasse 20
90403 Nuremberg, Germany
christian.pescher@fau.de
+49 911 5302 854

Gerard J. Tellis
Director of the Center for Global Innovation
Neely Chair of American Enterprise
University of Southern California
3670 Trousdale Pkwy
Los Angeles, CA 90089, USA
tellis@marshall.usc.edu

Johann Füller
Leopold-Franzens-University Innsbruck
Innrain 52
6020 Innsbruck, Austria
Johann.Fueller@uibk.ac.at

Can AI Help in Crowdsourcing?
Testing Alternate Algorithms for Idea Screening in Crowdsourcing Contests

Abstract

Crowdsourcing, while a boon to ideation, generates thousands of ideas. Screening these ideas to select a few winners is a major challenge because of the limited number, expertise, objectivity, and attention of judges. This paper compares original and extended versions of three recently published theory-based algorithms from marketing to evaluate ideas in crowdsourcing contests: Word Colocation, Content Atypicality, and Inspiration Redundancy. Each algorithm suggests predictors of winning ideas. The authors extend these predictors using two methods for searching parsimonious predictors: least average shrinkage and selection operator (LASSO) and K-sparse Exhaustive Search, for $K \leq 5$. The authors test the algorithms in-sample and out-of-sample on 21 different real-world crowdsourcing contests conducted for large firms. The standard provided by management is "drop the worst 25% of ideas without sacrificing more than 15% of good ideas," as ranked by experts. Results are the following. First, of the three original algorithms, Inspiration Redundancy performs best out-of-sample, but fails to meet the 15% threshold. Second, for two of the three algorithms, the extended versions outperform the original. In particular, Topic Overlap Atypicality, a new measure, emerges as the most robust predictor. Third, when the best versions of the algorithms are used, all three contribute to the important out-of-sample prediction accuracy. Fourth, using extended versions of all three algorithms, we are able to meet Hyve's threshold.

Keywords: Creativity, Prototypicality, Crowdsourcing, statistical learning, natural language processing, and network theory.

1. Introduction

Idea generation (ideation) and idea screening are important steps in the new product development process (Toubia and Flores 2007). They belong to the "fuzzy front end" which is a key leverage point of the firm (Dahan and Hauser 2001; Eling, Griffin and Langerak 2014). Crowdsourcing is a low cost way to generate many ideas (Howe 2008). The use of crowdsourcing revolutionizes ideation because it taps diverse information sources and perspectives and generates thousands of potential ideas or solutions at low cost. Commercial crowdsourcing platforms offer ideation-related services to various clients, typically acting as intermediaries between the crowd and a company seeking to produce ideas (Allen, Chandrasekaran and Basuroy 2017; Luo and Toubia 2015).

The main benefit of crowdsourcing ideas, many ideas, also presents the key challenge: screening ideas to identify the best. This screening process can be performed by users, i.e., contest participants or Amazon Mechanical Turk Workers, or by expert judges. Relying solely on the evaluations of contest participants for the evaluation of ideas may be problematic because contest participants may act strategically. For example, participants may vote down good ideas that compete with their own. Using Amazon Mechanical Turk Workers can be problematic because of the potentially low expertise of such workers for this highly specialized task. The use of expert judges to screen ideas is challenging because of their limited cognitive capability (Toubia 2006), attention span, expertise, or objectivity. A solution to the latter problem may be to let numerous experts each screen a small set of ideas. Prior research on consumers indicates that such a division of work may lead to the selection of good ideas (Toubia and Flores 2007). However, there may not be enough expert judges that have the knowledge necessary to assess ideas reliably.

A solution to this problem of screening may be to use theory-based algorithms as early attempt to use artificial intelligence (AI) in ideation. The biggest practical benefit would be screening out bad ideas at the very start of the idea evaluation process, when the ideation "funnel" is widest. If the theory-based algorithms manage to screen out bad ideas, humans would still be needed for both, ideation and idea selection. The elimination of bad ideas would lower the judges' cognitive load; this time saving allows

them to focus on the good ideas and – potentially – to take better decisions. So the first step in using AI for ideation is to determine if AI can screen out bad ideas. Once that goal is reached, AI can start to address problems in narrower parts of the funnel. If AI could select the best out of many new ideas, humans would become less important to the process of idea selection, but they would still be needed for ideation. The ultimate goal for AI in ideation would be to develop ideas which were universally recognized as better. If AI managed to do that, humans would become less critical to the ideation and idea selection processes. The broad goal of this paper is to attempt to take the first step of AI in ideation – screen out bad ideas in crowdsourcing.

To do so, this paper compares three[1] recently published original and extended versions of theory-based algorithms published in the marketing literature: Word Colocation, Content Atypicality, and Inspiration Redundancy. Henceforth in this paper, the word algorithm refers to only the three original and modified theory-based algorithms tested for this study. In this paragraph, we present the intuitions of these approaches; we elaborate on the theoretical backgrounds of the respective approaches in the theory section. The first algorithm, developed by Toubia and Netzer (2017), is what we call Idea's Word Colocation. The intuition of this approach is that good ideas balance novelty and familiarity. The second algorithm is Topic Atypicality used in Berger and Packard (2018). The intuition of this algorithm is that ideas are better if they are different from other ideas submitted to the contest. The third algorithm is the Ideator's Inspiration Redundancy used in Stephen, Zubcsek and Goldenberg (2016). The intuition of this approach is that ideators who possess access to diverse inspiration, i.e., they possess contacts to other ideators who do not communicate with each other, tend to submit better ideas.

Each algorithm suggests predictors of winning ideas. Because the algorithms have different theoretical foundations, the predictors may complement each other. We can test if they are complements or subsititutes. We extend these predictors using two different methods to identify the most parsimonious

[1] We are not aware of any other theory-based algorithms for evaluating ideas.

set of predictors: least average shrinkage and selection operator (LASSO) and K-sparse Exhaustive Search, for $K \leq 5$.

The authors of these three alorightms all evaluate their proposed algorithms using in-sample performance (fit). In-sample fitting may exploit sample idiosyncrasies rather than underlying patterns of creativity. It remains unclear whether the algorithms also do well in out-of-sample prediction. High out-of-sample predictive power is a prerequisite to completely or partially automate idea screening. To the best of our knwoeldge, no study has done the following: 1) compared the performance of these alternative algorithms in assessing ideas in the same context cordsouring consisting of 21 contest, 2) tried to improve upon them with new parsimonious predictors, and 3) analyzed whether or not these algorithm are substitutes or complements, and 4) assessed how the algorithms performed in in-sample versus out of sample prediction.

Each of the three algorithms has made a theoretical contribution in its original context of application. Yet, a) a rigorous performance assessment of the algorithms on a dataset in another context and b) an assessment whether theory-based improvements are possible may provide insights about the boader practical applicability of each algorithm and the respective underlying theory. Obtaining such a result may be a first important step in assessing the applicability of AI in ideation. Additionally, since each of the algorithms has an underlying theory, a rigorous text may indicate which of the theoretical arguments is strongest. However, it may as well be that the algorithms complement each other; if that were the case, our study may be a first step towards a unified theory for the complex phenomenon of creativity.

The specific goals of this paper are the following: 1) Compare the performance of the three algorithms from the marketing literature using in- and out-of-sample predictions. 2) Extend these algorithms based on theory, and compare the performance of original and extended algorithms, 3) do so using in- and out-of-sample predictions, and 4) assess whether the algorithms are rival or complementary. 5) Test whether any of these three algorithms, or a combination of them, can assist human experts in evaluating ideas in crowdsourcing contests, by at least screening out the large number of bad ideas.

To assess the performance of these algorithms, we use 21 crowdsourcing contests conducted for mainly large corporations and run by Hyve, a European innovation company. These contest can be grouped into three samples. Sample A consists predominantly of small contests, Sample B consists predominantly of large contests, Sample C combines Samples A and B. The standard provided by the Hyve's manager of the contests for a potential use of the algorithms is that it would already be helpful to screen out the 25% worst ideas without sacrificing more than 15% of "good" ideas, where "good" is equivalent to making the shortlist, which is the set of ideas selected by company experts for further consideration. The shortlist is determined by company experts, who possess a profound knowledge on the subject and who are responsible for the implementation of the winning ideas and who, thus, have a profound self-interest in selecting the best ideas. We obtained considerable data from Hyve about these contests for research and not consulting purposes. However, to preserve the confidentially of their clients and ideators, Hyve was unable to provide complete details of all the contests.

The results of our research to meet Hyve's standard are the following. First, acknowledging that each original algorithm provided unique theoretical and methodological value, Inspiration Redundancy performs best out-of-sample, but fails to meet the 15% threshold. Second, for two of the three algorithms, the extended versions outperform the original. In particular, Topic Overlap Atypicality, our new measure which is a measure of the number of topics an idea has in common with the rest of the crowd, emerges as the most robust predictor. Third, in terms of ***out-of-sample prediction accuracy,*** when the best-performing versions of the algorithms are used, all three contribute to this criterion.

We make the following contributions to literature and practice.

First, we show the importance of testing algorithms outside of the estimation or training sample, while prior work has focused on in-sample prediction. Our results show that in-sample predictions typically lead to better results than out-of-sample predictions across measures. Using out-of-sample prediction is a more rigorous test because in-sample predictability may exploit idiosyncrasies of the data. The finding that Inspiration Redundancy, which is an ideator-based approach, outperforms the two text-

mining based approaches out-of-sample, implies that an ideator's traits are particularly important in ideation.

Second, a) We generate predictions based on a new word colocation network generated from Google Patents. The network represents the wisdom of experts and is an extension of Toubia and Netzer's (2017) wisdom of the crowd based on Google Search. b) We develop a new approach for measuring Topic Atypicality, called Topic Overlap Atypicality, whose key component is the number of topics an idea has in common with a reference corpus (divided by the number it could have in common). c) We apply Burt's (2004) Constraints measure to idea screening as an extension of Stephen, Zubcsek and Goldenberg's (2016) Inspiration Redundancy.

Third, our result that extended versions tend to outperform the original algorithms shows that one can reach important gains in the particularly important out-of-sample prediction accuracy either by developing new algorithms or by carefully expanding existing approaches. Our results are that when the best versions of the algorithms are used, all three contribute to prediction accuracy. Currently, the best algorithms in marketing can only focus on one or a few important dimensions, whereas the human mind is multi-faceted and the setting of innovation is very complex, which AI currently fails to fully capture.

Fourth, at the current stage of assessing ideas, AI cannot fully or even partially substitute for human experts. However, AI can assist humans, in our case Hyve's managers, by screening out bad ideas (Level 1), thereby reducing human's cognitive load and help them focus on the good ideas. Thus, the current study is a first steph to narrow the ideation funnel early on. Yet, as research advances and as additional algorithms bring in unique and helpful perspectives, research may develop algorithms to select the best ideas (Level 2) or even to automatically generate outstanding ideas (Level 3).

2. Literature

Our work is closely related to the literature on crowdsourcing (Allen, Chandrasekaran, and Basuroy 2018; Stephen, Zubcsek and Goldenberg 2016; Toubia and Netzer 2017), and in particular the screening process for handling these large quantities. The three algorithms of interest in this paper are: Idea's Word Colocation (Toubia and Netzer 2017), Topic Atypicality (Berger and Packard, 2018), and Ideator's

Inspiration Redundancy (Stephen, Zubcsek and Goldenberg 2016). We review the theoretical background of these approaches to guide our research.

2.1 Idea's Word Colocation

Novelty versus Familiarity. Many new ideas and concepts are the outcome of a process of combination and reorganization of existing ideas and concepts (Mobley et al. 1992). Creativity consists of reassembling elements from existing knowledge bases in a novel fashion (Dahl and Moreau 2002). Thereby, within an idea, a moderate level of incongruity between the concepts that make up the idea can be beneficial (Finke et al. 1992). Research has identified a theory about optimal levels of incongruity: the concept of familiarity versus novelty (Toubia and Netzer 2017) draws on the cognitive perspective of creativity (Dahl, Chattopadhyay, and Gorn 1999; Goldenberg and Mazursky 2002), which asserts that evaluators rely on information stored in their memories to judge ideas (Toubia and Netzer 2017). This stored information consists of preinventive structures, i.e., symbolic patterns, exemplars, mental models, or unique verbal combinations that exist in the mind of the evaluator (Burroughs et al. 2008). When individuals perceive stimuli related to their knowledge, the stimuli activate their domain-specific schemas (Bilalić et al. 2008). Because of schema activation Toubia and Netzer (2017) argue that a) if an idea is too novel, its evaluation takes place largely in a vacuum and the evaluator will not know how to judge it; if an idea is too familiar, it seems to be rather incremental – or not new or interesting at all. Thus, experts rank highest those ideas that optimally balance novelty and familiarity based on their current state of knowledge.

Wisdom of Crowds. With modern text-mining methods, word colocation networks can be constructed in seconds, depending on the source2. In a word colocation network, the vertices are word stems or word lemmas and edges indicate co-occurrence. Word stems or lemmas that appear together more frequently have higher weights and are therefore "closer" to each other (Netzer et al. 2012). Toubia and Netzer (2017) use the group of edge weights in an idea to measure its balance of "novelty" compared

[2] Typically the time constraint comes from downloading text, as opposed to processing it.

to “familiarity”. A key point in the paper is that it is not the words that directly determine novelty or familiarity, but instead, the combinations of words within an idea. Thus, for Toubia and Netzer an idea is novel to the degree it contains words that typically do not appear together. It is familiar to the degree that it contains words that frequently appear together.

Toubia and Netzer (2017) use Google Search results to construct a word colocation network. This approach implicitly assumes that Google Search results represent the “wisdom of crowds” because they are rank-ordered based on the activity of a very large group of users, so that a word colocation network computed from averaging over high ranking (roughly top 50) results is thought to contain a desirable balance between “novelty” and “familiarity.”

Wisdom-of-Experts. Some literature reports that consumers’ evaluations predict market success as well as or better than experts (Kornish and Ulrich 2014). Yet, that literature focuses on incremental innovations like new designs, which can be introduced into the market shortly after the contest. In contrast, the evaluation of the idea quality for radical innovations or complex products which first need to be developed, tested, and introduced into the market often has to take place years before their market introduction. During that time period, consumer preferences can change substantially. In such settings, companies rely on the knowledge and experience of their experts to judge idea quality. In addition, experts are more capable of determining implementation feasibility compared to consumers. Thus, to successfully screen ideas in these situations, experts’ knowledge is of primary importance.

Knowledge is stored in semantic structures. Research across disciplines suggests that these semantic structures differ between experts and non-experts. Specifically, experts use more abstract thinking and more technical language. Researchers present such examples in medicine (van de Wiel et al. 2000), computer science (Hughes and Parkes 2003), or even among children in their areas of interest (Gobbo and Chi 1986). For example, articles from the popular press are likely to use simple language, easily understandable to most readers. On the other hand, scientific publications are typically complex and use many technical terms. Only a few hundred experts (sometimes even less) can fully understand them.

As stated earlier, Google's Search result are thought to represent the wisdom-of-crowds and thus are likely related to customer predictions (Toubia and Netzer 2017). However, in predictions (van Bruggen, Lilien and Kacker 2002), datamining (Hill and Ready-Campbell 2011), and social media recommender systems (Amatriain et al. 2009), research shows that a selected subsample of experts tend to outperform the entire crowd. In the case of Wikipedia, harnessing the wisdom of a few self-selected experts has proven useful (Kittur and Kraut 2008).Thus, relying on experts may complement relying on crowds.

Our extended methods include consider these findings and, thus, rely on the knowledge of experts. For this purpose, we use corpus of Google Patents to construct the reference networks. Google Patents captures more technical vocabulary and relationships between words than does Google Search. Because Google Patents depends on the knowledge of a comparably small group of knowledgeable contributors, particularly if one compares it with Google, we label the reference corpus the "wisdom of experts."

2.2 Typicality versus Atypicality of idea

The internet contains a large amount of freely accessible text. One important contribution of Toubia and Netzer (2017) is to show how publicly available data can be used to potentially automation idea evaluation. Their measures – whether they represent the wisdom-of-crowds or the wisdom-of-experts – access "global" information (i.e., information not specific to the evaluation context) to assess the quality of ideas.

A growing body of research suggests that creativity also has a context-specific dimension (e.g., Baer 1998). For our setting, one type of context-specific information for use in evaluating a given idea is the group of other ideas submitted to the same contest. The question is whether atypicality is positively or negatively related to idea quality. The literature provides evidence for both. In music, a creative endeavor, songs with atypical lyrics, i.e., lyrics, which diverge topically from a genre average, are more likely to become successful (Berger and Packard 2018). The intuition behind that finding is that if a song says something unconventional in its genre, consumers are more likely to pay attention to it.

In contrast, in the ideation-related setting of parallel search, i.e., a setting in which ideators cannot see each other's ideas, typical ideas tend to possess a high value (Kornish and Ulrich 2011). Typical ideas are similar to all other ideas in the contest. The appealing intuition behind this insight is that if many people come up with similar ideas independently from one another, a latent consumer need can be addressed to create value. However, this insight was generated letting students come up with ideas to rather incremental tasks, e.g., physical products for college student market with a retail price of less than $50, and then to use a peer-evaluation mechanism to judge idea quality. The similarity of ideas was assessed by grouping the ideas according to the respondents' own notion of similarity and then comparing the similarity between groups of ideas. However, it remains unclear whether this finding generalizes to real-world contests with idea visibility evaluated by company experts.

2.3 Ideator's Inspiration Redundancy

The key idea of interconnectivity of ideators is not to analyze the ideas' text, but to consider the inspirations of the ideator. If ideas submitted to crowdsourcing contests are visible to other participants, it can inspire and potentially influence them (Wooten and Ulrich 2019). The network structure that surrounds ideators can provide information about the redundancy of their inspirations – which in turn influences the quality of the ideas they submit (Stephen, Zubcsek and Goldenberg 2016). Three literature streams provide input for this concept: (1) the wisdom-of-crowds literature argues that crowds can be less wise if interdependencies increase (e.g., Pentland 2013), (2) organizational sociology shows that interconnectivity can affect performance in problem-solving including creative and innovative thinking (e.g., Burt 2004), and (3) brainstorming literature illustrates that individuals can outperform groups in ideation (e.g., Lamm and Trommsdorff 2006).

If the ideator's network neighbors are not connected to one another, ideators receive independent inspirations. In contrast, if the network neighbors are connected, they also influence each other and ideators receive similar, redundant inspirations (Burt 2004). Stephen, Zubcsek and Goldenberg (2016) name the following reasons why higher redundancy leads to lower quality ideas: (1) a decreasing size of the set of neighbors' ideas which serve as inspirations when ideating may lead to decreasing

innovativeness, (2) idea redundancy could stifle individual innovativeness, because it interferes with psychological mechanisms like fixation (Bayus 2013) involved in processing others' ideas, and (3) the recurrence of an idea can operate as a proof signal. These mechanisms may lead to similar ideas and to a decrease in variance of idea quality in the contest. However, if the mean idea quality is constant, a high variance of idea quality is desirable, because it increases the likelihood of finding one or a few outstanding ideas (Terwiesch and Ulrich 2009). Thus, increasing interconnectivity can be problematic in ideation.

3. Empirical Approach

Here we describe how we operationalize the theories described above. Methodologically, two of the algorithms are text-mining approaches and the third one is a social network approach. We describe the methods of each of the three algorithms, along with our extensions.

3.1 Text Mining Methodology

Text mining (Berger et al. 2020; Netzer et al. 2012; Netzer, Lemaire and Herzenstein 2019) has become increasingly popular as a research tool since it helps to detect patterns in large unstructured text corpora, which can be used to generate knowledge about consumers (Matz and Netzer 2017; Wedel and Kannan 2016).

Text Pre-Processing. In each case, we prepare the text by eliminating "stopwords," or words which appear extremely commonly (e.g., "the" and "and"). We also run a process called lemmatization which stands in place of the traditional approach of stemming. While stemming simply truncates words in order to reduce duplicates of the same word, lemmatization attempts to remove inflectional endings and reduce words to a base dictionary word. Words stems are not always words, while word lemmas always are. We chose lemmatization over stemming mainly for convenience in working with the text, since it is easier to make sense of word lemmas than word stems. The impact on downstream performance is likely to be insignificantly small, but lemmatization may be very slightly better (Balakrisnan and Lloyd-Yemoh, 2014).

Idea's Word Colocation. We apply and extend the methods from Toubia and Netzer (2017). This approach uses a number of metrics computed from the words used to describe an idea. The main building blocks for these various metrics are word frequencies and Jaccard indices. Both word frequencies and Jaccard indices require a reference corpus and a word colocation network in order to be computed. As mentioned above, Toubia and Netzer (2017) introduced a novel reference corpus consisting of the first 50 Google results when the ideation topic is entered as a search term. The word frequencies are simply the number of times each word in the idea appears in the reference corpus. The Jaccard index, computed on a word *pair*, is the intersection over the union of documents containing the respective words in the pair. Here, the word "document" refers generally to a body of text. In practice, documents are specified by the researcher (e.g., a Google Search result or a Google Patent result). If i and j are words and D_i and D_j are sets of documents containing them, respectively, then the Jaccard index between i and j is:

$$J(i,j) = \frac{D_j \cap D_i}{D_j \cup D_i}$$

For example, if the documents are "one two three", "one two" and "one", then the Jaccard index between the words "one" and "three" is 1/3, since one document contains both while all three contain at least one of the two. With the Jaccard indices and the word frequencies from each idea, Toubia and Netzer (2017) construct a number of metrics: the average, max and min word frequency of an idea, the average max, and min Jaccard indices from an idea, the coefficient of variation of word frequencies, and the coefficient of variation of Jaccard indices. A key metric is the Kolmogorov-Smirnov (KS) metric, which is created by first computing Empirical Cumulative Distribution Functions (ECDFs) for each document of a reference corpus, taking the average of those results, and then comparing the ECDF of a given idea to the average using the KS metric. The KS metric is simply the maximum absolute difference between two vectors. Toubia and Netzer show that the KS metric relates negatively to idea quality, even after controlling for other word-derived metrics: the KS metric is their measure to balance novelty with familiarity.

We refer to this approach as the Idea's Word Colocation approach because Jaccard indices measure how often word pairs are collocated. We substantially extend Toubia and Netzer's (2017) approach by comparing several reference corpora; this idea is based on findings from theory that the wisdom-of-experts can outperform the wisdom-of-crowds. Toubia and Netzer (2017) show that their KS result holds when the reference corpus is a set of Google results related to the idea generation topic. We test different reference corpus. In addition to Google search results we use Google Patents. The Google Patents corpus is generated the same way as the Google Search corpus, but with Google Patents as the search engine.

Using each reference corpus, we compute the metrics used in Toubia and Netzer (2017): the mean, min, max, and coefficient of variation for both Jaccard indices and word frequencies of each idea, and the KS metric for each idea.

Topic Atypicality. This metric is developed in the spirit of Berger and Packard (2018). Their application context is further removed from ours compared to the other two: instead of creativity in an ideation context, they study creativity in a music setting. The dependent variable is song popularity. Berger and Packard use Latent Dirichlet Allocation (LDA) to generate a topic model from a corpus of song lyrics. LDA "is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities." (Blei et al. 2003). In other words, LDA views topics as having a relationship with individual words: some words are likely or unlikely to appear for certain topics. A common representation for LDA model outputs is a set of linear equations, one for each topic, where the equation terms represent words. Each word term has a coefficient indicating the importance of that word to the topic. These equations can be inputs into operations. Specifically, a topic score can be computed for a document, by applying the model weights against each word in the document. Thinking of the topics as coordinates, each document can be imagined to have a 'location' somewhere in topic space. In our case, we create an LDA model for the entire set of ideas in a contest and

compute its topic location by applying the LDA weights to its vocabulary. Then for each idea we compute the distance between it and the location of the overall corpus in topic space.

Berger and Packard use Ireland and Pennebaker's (2010) style matching equation to measure distance, which is sensitive to the degree of deviation within each topic. For a topic *t*, idea *i*, with reference corpus *C*, the style-matching atypicality $A_{SM}(C, i)$ is given by

$$A_{SM}(C, t, i) = \frac{1}{1 - |t_c - t_i|/(t_c + t_i + .001)}$$

The total Style Matching Atypicality, $A_{SM}(C, i)$ is calculated by taking the average of each $A_{SM}(C, t, i)$ over topics *t*.

In addition to the style matching atypicality, we also develop a new metric, Topic Overlap Atypicality A_J, which ignores the degree of deviation for an individual topic, and simply indicates presence or absence of deviation. Specifically, it counts the number of topics in common between the reference corpus *C* and each idea, and divides that by the total number of topics in either one, then subtracts this quotient from one. Mathematically, if T_C is the set of topics in C and T_i the set of topics in *i*, then the modified atypicality is given by:

$$A_J(C, i) = 1 - \frac{T_C \cap T_i}{T_C \cup T_i}$$

which is simply the Jaccard distance between the topic sets. Because of its construction, we expect style matching atypicality, $A_{SM}(C, i)$, to perform well if the degree of departure from average topics is informative. On the contrary, since style matching atypicality is an average, it tends to be sensitive or even be misguided by noisy outliers. However, because Topic Overlap Atypicality, $A_J(C, i)$ is a count variable, we expect it to perform well when the incidence of topic departure matters, but when the distance is less informative. Since the method does not get biased by strong outliers, it tends to be more robust in the context of idea generation. This is because idea generation creates a large number of diverse ideas. Some of these ideas will naturally fall far away, topically, from the "average" idea. A metric that is less influenced by these outliers will more strongly reflect the topics which appear more frequently.

3.2 Network Methodology

Methods previously used to study social networks have recently been employed for studying creativity and ideation. (Stephen, Zubcsek and Goldenberg 2016; Godart and Galunic, 2019). We explain the two methods we tested.

Ideator's Inspiration Redundancy. This measure comes from Stephen, Zubcsek and Goldenberg (2016). For each ideator, we create a network where nodes are ideators and links are comments. If one ideator has made a comment on another ideator's idea, those two ideators are connected. Otherwise, they are not connected. Using this network, we compute the Clustering Coefficient (Watts and Strogatz 1998) for each ideator. The clustering coefficient for a node *i* in an undirected graph is the number of edges among *i*'s neighbors that exist divided by the number that could exist. If N_i is the "neighborhood" of *i, j* and *k* are neighbors of *i*, e_{jk} is an edge between *j* and *k*, v_j and v_k are nodes for *j* and *k,* and node *i* has k_i total neighbors, the clustering coefficient CC_i is:

$$CC(i, N_i) = \frac{2|\{e_{jk}: v_j, v_k \in N_i\}|}{k_i(k_i - 1)}$$

Since the unit of analysis is ideas, we then simply apply the Clustering Coefficient of each ideator to all of his or her ideas. This clustering coefficient is computed for the group of comments at the very end of the contest. We refer to this method as the Ideator's Inspiration Redundancy because the Clustering Coefficient measures the degree of connectivity of each ideator's sub-community. Since comments reflect the information present in the network surrounding an ideator, the Clustering Coefficient measures the diversity of information around the ideator. Stephen, Zubcsek and Goldenberg (2016) showed in their context that higher Clustering Coefficients related negatively to idea innovativeness as judged by consumers The reason is that high clustering indicates that the inspiration sources in that area of the network are less diverse.

In addition to the clustering coefficient, we use Constraints measure (Burt 2009). This measure has recently been used to measure the popularity of cultural elements in fashion (Godart and Galunic 2019). Whereas the clustering coefficient only considers information from the direct neighbors, the

constraints measure also considers second order neighbors (i.e., neighbors of neighbors). The constraints measure reflects the value of a particular node as a "bridge" between nodes which are otherwise rarely bridged. In the context of creativity, connections between concepts, people or facts which are not typically connected have higher potential for novelty.

Let a_{ij} indicate the existence of a link between *i* and *j*, so it equals 1 if that is the case and 0 otherwise. Then define $p_{ij} = a_{ij}/N_i$. Then Constraints measure is

$$C(i, N_i) = \sum_{j \neq i} (p_{ij} + \sum_{k \neq i; k \neq j} p_{ik} p_{kj})^2$$

This measure sums across neighbors of a node *i*. For each neighbor *j*, the constraints measure for node *i* is larger if *i* and *j* have many neighbors in common. The measure is also larger if *i* has fewer connections. We expect this measure to be negatively related to idea quality, as larger values indicate that node *i* has fewer connections and connects fewer groups of otherwise unconnected nodes.

For consistency with Stephen, Zubcsek and Goldenberg (2016), we also include the Degree of each idea in both the original and new version of the Inspiration Redundancy algorithm.

3.3 Data Sources

We compare the predictive accuracy of the Idea's Word Colocation, Topic Atypicality, and Ideator's Inspiration Redundancy algorithms on 21 different crowdsourcing contests conducted for large corporate clients, on the crowdsourcing platform of Hyve, an innovation company. We pool data from these 21 contests into three samples: Sample A contains eight small contests, Sample B contains eight larger contests, and Sample C combines Sample A and Sample B. We use different samples in order to understand how the methods work. The contests are grouped to samples according to the design of the idea submission website, i.e., the contests from sample A typically had a different design than the contests from sample B. The other differences between Samples A and B are incidental. These contests were conducted for client firms such as Frankfurt Airport, Lufthansa, MasterCard, Deloitte and Telekom, Vodafone, Zeiss, Volkswagen, and DHL. Typically, both, Hyve and its client, recruit ideators by public announcements and by contacting ideators from past contests. The contest usually answers one specific

question and runs for a limited time between 30 and 80 days; the deadline is communicated before the contest starts. All contests are idea generation contests which search for innovative ideas about future products, services, or business models. We only use data from contests in which the ideas are verbally described. This excludes another popular form of contests, design contests (cf. Allen, Chandrasekaran and Basuroy 2018). The contests run on the same platform and offer social networking functionalities: ideators can explore, evaluate, and comment on the ideas of others. However, there is no formal collaboration system. The idea evaluation occurs in three stages: expert ratings of all ideas, experts select shortlist of ideas to be presented to the jury, jury's selection of winners.

Expert Ratings. Second, after the end of the contest, experts on the topic, usually from the client company, evaluate all ideas on a 5-point scale. The clients gave experts specific criteria to evaluate ideas according to contest-specific criteria. Unfortunately, we do not have full access to all these criteria. So, we have the outcomes (ratings) of the ideas but not the criteria that led to them.

Experts' Shortlist of Ideas. Third, based on their own evaluations, the experts build up a shortlist of 12 to 30 ideas. Usually, the experts' top-rated ideas make the shortlist. However, if one expert likes a specific idea particularly well, he or she can discuss it with other experts; if the other experts do not oppose, such an idea can additionally make the shortlist. All shortlisted ideas usually receive at least a small prize or some formal recognition.

Jury's Selection of Winners. Fourth, the shortlist of ideas as well as a contest overview is presented to a jury of 5 to 10 members, which usually consists of client's top executives, experienced innovators, professors, or consultants on the topic of the contest. The juries select the winners in two ways. Either the jury members individually vote on the ideas using a scorecard tailored to specific innovation criteria and simple aggregation determines the winners or the jury jointly selects the winners in a discussion. In addition to the ideas proposed by the experts, jury members have the option to pick and evaluate any idea from the contest. Or, the juries determine the winners in discussion sessions.

3.4 Measures

This section provides a description of the dependent variable and independent variables.

Dependent variable: success. The dependent variable is success. We consider an idea a success (1) if it makes the experts' shortlist. We choose this success measure for three reasons:

First, every shortlisted idea receives some prize or formal recognition, which means, a reward. We consider receiving rewards to be a measure of success.

Second, in private discussions, managers of Hyve state that they have a high degree of confidence in the shortlist, but not in the winner.

Third, the juries select very few winners relative to the thousands of ideas. Ten contests had a total of 12 winners, some contests had more than one winner. Such a rare event in the dependent variable results in a low variance – which models typically cannot capture in a meaningful way. Contrary, the experts' shortlists consist of up to 30 ideas and, thus, are a richer success measure than the jury's winners. So, we code shortlisted ideas as 1, other ideas as 0.

Independent variables. Table 1 lists the independent variables, which we explain below.

Insert Table 1 here

The three algorithms described above produce the independent variables. The Idea's Word Colocation method produces 9 variables for each reference corpus (max, min, mean, coefficient of variation for Jaccard indices and node frequencies, along with the Kolmogorov-Smirnov metric). We use two reference corpora: Google Search and Google Patents. The Topic Atypicality method produces two variables: Style Matching Atypicality and Topic Overlap Atypicality. The Ideator Inspiration Redundancy method produces three variables: ideator degree and clustering coefficient and Burt's constraints measure.

Our goal is to distinguish among ideas within the same contest. Any variables which are constant for all ideas in a contest are absorbed into the model's intercept. For this reason, we don't explicitly include conventional econometric "controls" such as contest size, date, industry, etc. Because our analysis is intra-contest, control for those variables are not necessary.

3.5 Selection Techniques

To find the most robust predictor variables, use two variable selection techniques. The first is the least average shrinkage and selection operator (Lasso). The second is K-sparse Exhaustive Search. We explain both techniques here.

Lasso. For Lasso, we place each predictor variable into a logistic regression[3]. The dependent variable in this regression is an indicator of whether an idea is on the 'shortlist':

$$y_i = \begin{cases} 1 \text{ if idea } i \text{ is shortlisted} \\ 0 \text{ otherwise.} \end{cases}$$

We use a logistic regression to accommodate this binary dependent variable. To produce predictions, we sort the ideas by the model's estimated probability of being shortlisted, and take the top *n* ideas, where *n* is the number of shortlisted ideas in the actual contest.

Lasso regression is a regularized version of standard regression (Tibshirani, 1996, Sun et al., 2019). Regularization is an explicit bias, built into a predictive model to move it in the direction of some desirable property. In the case of lasso, the model has a bias toward parsimony: during estimation the model attempts to assign coefficients of zero to variables with lower weights. Lasso is like a conventional logistic regression, except there is no "error" term associated with a Lasso regression and the sum of the coefficients in a Lasso regression is "penalized" according to a tuning parameter (λ.) It is useful to think of Lasso as a conventional regression but with a coefficient "budget." The tuning parameter controls the size of the budget. The parameter λ is referred to as a tuning parameter because it is "tuned" using the training set. We follow a typical "statistical learning" (Friedman et al., 2001) workflow by tuning the parameter on a training set, and then by assessing predictive accuracy on a test set (not included in the training set).

We use cross-validation to tune λ. In cross-validation, the sample is split into subsets, the model is trained on all but one of those sets, and the remaining set is used as a holdout to compute an accuracy

[3] Since binary logistic regression treats outcomes as independent, it would inappropriate if the number of shortlisted ideas were constrained to a certain number. This is not the case in our contests. There are no cases where one idea was no shortlisted because too many ideas already had been. In one contest, nearly one-third of all ideas were shortlisted.

metric. This is repeated so that each of the subsets takes a turn being the holdout sample, and the accuracy metric is averaged across the repetitions (Stone, 1974). Cross-validation is typically used to avoid overfitting in machine learning models (Friedman et al., 2001). In a cross validated Lasso model, the variables which receive coefficients of zero tend to be those which are less robust predictors (Abadie and Kasy 2020). Lasso can be used as a variable selection technique, as long as the interpretation is approached with care.[4] For our cross validation, we split the sample into the number of contests, and each subset is simply the group of ideas for that contest.

Lasso is often used because of its robust performance across many settings (Abadie and Kasy 2020). This makes it appealing in our case since our dataset includes contests about several different topics, ranging from beer festival moments to plastic extrusion machines. In order to be useful for our task, we must have predictor variables which are robust across contests which have varying topics and distinct judging panels.

K-Sparse Exhaustive Search. K-Sparse Exhaustive Search is a very simple technique for selecting variables. First, the researcher selects *K*, which is the maximum number of variables in a combination. Then, each possible combination of *K* variables is used in the prediction task, and an accuracy metric is computed. We provide details of our accuracy metric below. We use *K*-sparse Exhaustive Search to provide a "second opinion" about robust predictors. *K*-sparse Exhaustive Search is more difficult to interpret than lasso. The two variable selection techniques complement one another: Lasso helps to obtain a narrow but interpretable view; *K*-sparse Exhaustive Search provides a wider view which may be hard to interpret.

3.6 Class Imbalance

In our task of predicting shortlisted ideas, we face a problem which is common in machine learning, known as "class imbalance." A class imbalance is a case where values taken by the left-hand side variable

[4] Like in a regular OLS regression, the coefficients in a lasso model do not indicate causality. In addition, lasso's weightings require the tuning of a regularization parameter. We use cross-validation for this tuning, which involves a degree of randomness, so that the coefficients can vary between estimations even on the same dataset, however, the variation is small.

are unevenly distributed. For example, the number of shortlisted ideas is outnumbered at a roughly 16 to 1 ratio by those which are not shortlisted. This imbalance gives the model less information about shortlisted ideas than those which are not shortlisted.

Statisticians have designed a number of methods to improve predication accuracy in such settings. We chose to use a procedure known as "SMOTE", which stands for Synthetic Minority Oversampling TEchnique (Chawla et al., 2002) "The minority class is over sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the *k* minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the *k* nearest neighbors are randomly chosen." In some sense, this technique "fills in" some of the empty spaces between minority class observations. When we applied this technique, hold out performance improved notably.[5]

4. Results

We present results by sample. First, for each sample, we present what we call a "sacrifice rate" for each algorithm or combination of algorithms. The sacrifice rate is defined as the percentage of shortlisted ideas removed ("sacrificed") by eliminating the ideas which the model predicts to be in the worst quartile. Recall that shortlisted ideas are the "best" 12 to 15 of all ideas, as judged by company experts; the "best" ideas, typically, are presented to the jury, which then selects the winner. We use sacrifice rate as our metric because we face a unique prediction task: we must judge our accuracy with reference to the standard given by Hyve. This is to remove the worst 25% of ideas without removing more than 15% of shortlisted ideas. Thus, the error rate in the worst quartile of ideas is most important. An ideal model would have a sacrifice rate of 0%, so a quarter of ideas could be removed without sacrificing any shortlisted ideas.

[5] After applying SMOTE to the dataset, many of the variables are correlated, and this can often lead to poor variable selection by Lasso. For this reason, we select variables before applying SMOTE. Then we used the preselected variables on the modified dataset for final estimation results.

We calculate the in-sample sacrifice rate by grouping every idea and every predicted probability from every contest into a single dataset. Then we eliminate the worst 25% from this aggregated dataset and compute the number of shortlisted ideas sacrificed. The same is done for out-of-sample sacrifice rates.

The in-sample sacrifice table measures fit, as opposed to predictive accuracy, since the sample used to compute accuracy metrics is the same as the sample used to train the model. Out-of-sample sacrifice rates are computed by holding out each contest as a test sample in turn, and using the model to assign probability of shortlisting to each idea. Then we place all the ideas and predictions into one large dataset, sort it by shortlisting probability, and calculate the sacrifice rate on the bottom quartile. Last, we present the results of *K*-sparse Exhaustive Search, which take the form of a table of the top 10 variable combinations.

We do not use common standard performance metrics in statistical learning which include the Receiver Operator Curve (ROC) and Mean Squared Error (MSE) (Hanley and McNeal 1982; Hastie et al., 2001), because both give equal weight to classification errors across all observations, whereas our prediction task focuses only on false negatives among shortlisted ideas. However, for comparison, we provide F1 scores for each method in an appendix. The F1 score is the most appropriate conventional metric for our setting because it balances precision (true positives divided by all positives) against recall (true negatives divided by all negatives). The F1 tables generally show the same patterns as the sacrifice rate, but they are more difficult to interpret in the context of this task and they are also relatively insensitive across methods. These two characteristics support our decision to use a custom metric.

4.1 Sample A Results

Table 2 compares sacrifices across algorithms for Sample A. In general the algorithms perform worst in this sample. With a sacrifice rate of 13%, Topic Overlap Atypicality performs well. However, it is the only algorithm to meet the 15% threshold on Sample A. For this sample, each of the extended versions of the algorithms performs as well as or better than the corresponding original algorithms out-of-sample.

Table 3 shows the coefficients of the lasso logistic regression trained in-sample. Bold text indicates variables which are novel to this work. Four predictor variables receive non-zero weightings. The heaviest weighting is on Node Frequency Coefficient of Variation computed from the Patent reference corpus which has a standardized coefficient of 0.32. This result is not consistent across samples, suggesting that this variable is not robust. Topic Overlap Atypicality also receives a non-zero weight (-0.06). This is an important finding, because the remaining samples come to consistent results, which we will discuss below.

Table 4 presents the results of *K*-sparse Exhaustive Search, for *K*=5. The top 10 combinations are shown. In other words, the 10 best combinations of 5 or fewer variables are shown, where "best" is measured by the out-of-sample sacrifice rate.

Topic Overlap Atypicality is the only variable which appears in each of the top ten performing variable combinations. Clustering coefficient appears in eight of the top ten. Node Frequency Coefficient of Variation (Patent) appears in five. The most parsimonious combination in the top ten is Topic Overlap Atypicality and Max Node Frequency (Patent).

We visualize the sacrifice rates in Figure 1, which contains four plots: aggregated in and out-of-sample performance when all algorithms are used, compared with Topic Overlap Atypicality. We compare to Topic Overlap Atypicality because it is the best performing individual algorithm. In the figures labelled "All Algorithms, Lasso Selection," variables used are those selected by Lasso. The bottom two plots show sacrifice rates for each idea contest.

Each figure uses green to signify that the 15% threshold is met, yellow to mean the sacrifice rate is between 15% and 25% (random), and red to indicate sacrifice rates above 25%. It is immediately apparent that accuracy out-of-sample is poor when all algorithms are used. When exclusively Topic Overlap Atypicality, performance is better, but for a number of contests the sacrifice rate is still well above 15%.

4.2 Sample B Results

Table 5 compares sacrifices across algorithms for Sample B. In general the algorithms perform better in this sample relative to Sample A. In Sample B, the 15% threshold is met for three algorithms or algorithm combinations. Here, we see that the best performing algorithm out-of-sample is the modified version of Inspiration Redundancy, the Constraints Measure, with a sacrifice rate of 13%. The second best performing algorithms are Topic Overlap Atypicality and the combination of all algorithms (new and old), each of which have a sacrifice rate of 15%.

Table 6 shows the coefficients of the lasso logistic regression trained in-sample, on Sample B. Bold text indicates variables which are novel to this work. Nine predictor variables receive non-zero weightings. The heaviest weighting is on the Clustering Coefficient, with a standardized coefficient of -0.32. We note that a negative weight is consistent with the theory and result of Stephen, Zubcsek and Goldenberg (2016). Topic Overlap Atypicality also receives a non-zero weight (-0.21). This is consistent with the results from Sample A, where Topic Overlap Atypicality also received a non-zero, negative weight.

Table 7 shows the results of *K*-sparse Exhaustive Search, for *K*=5. The Constraints Measure appears each of the top ten for this sample. Average Node Frequency (Patent) appears in five, and Clustering Coefficient and Topic Overlap Atypicality appear in only three. The most parsimonious combination in the top ten is the Clustering Coefficient, Constraints Measure, and Average Node Frequency (Patent).

We visualize the sacrifice rates in Figure 2. The performance on Sample B is dramatically better than on Sample A. In general, this is likely to be due to two important factors: the average number of ideas per contest in Sample B is higher than in Sample A, and additionally the average number of shortlisted ideas per contest is larger in Sample B. In other words, the algorithms have more information per contest to learn from.

4.3 Sample C Results

We now turn to Sample C, which combines Sample A and Sample B. Table 8 shows the sacrifice rates across algorithms for Sample C. With a sacrifice rate of 14%, Topic Overlap Atypicality ties with the Combined Original and New.

Table 9 shows the non-zero Lasso weights across all variables. Topic Overlap Atypicality is the variable with the largest magnitude weight (-.17); the only other variables selected is Node Frequency Coefficient of Variation (Google), with a very small weight of 0.02. Topic Overlap Atypicality is the only algorithm to have a non-zero weight in all three samples. It also achieves equal or better accuracy than all the algorithms combined in each sample. Topic Overlap Atypicality only narrowly misses being the most accurate algorithm for each sample, only slightly exceeded in accuracy by the Constraints Measure in Sample B. Topic Overlap Atypicality is the only algorithm to meet the 15% threshold in every sample.

Table 10 shows the best variable combinations for Sample C from K-sparse Exhaustive Search, with K=5. Just as in Sample A, Topic Overlap Atypicality appears in each of the top ten. Notably, the Constraints Measure appears in 9 of the top 10 variable combinations that perform best, yet assigned a weight of zero by Lasso. This discrepancy is probably due to the mechanism for variable selection employed by Lasso, which is based on relative coefficient size. The reason for this result is that the Constraints Measure probably has a small enough coefficient that it was "crowded out" by other variables. Because K-sparse Exhaustive Search tries every combination, it reveals some combinations that Lasso would not try. However, trying every combination does not make K-sparse necessarily superior, because it may so doing leads to overfitting to this dataset. In contrast, Lasso is more robust because it focuses exclusively on predictors with larger effects sizes (Abadie and Kasy 2020). We use the two methods together in order to capture with one method what the other may have missed. We visualize the sacrifice rates in Figure 3. We see that the best algorithm meets the 15% threshold provided by Hyve (green) for 11 contests of 21.

4.4 Simulation to Construct a Benchmark

We compare this result, i.e., reaching the threshold for 11 out of 21 contests, to the performance of a random benchmark. For how many contests would a random guesser meet the threshold? Using the

number of shortlisted ideas and the sample size of each contest in our dataset we can compute the probability that a random guesser would meet the 15% threshold. The problem of randomly sampling from a set of objects of different types is a classic problem in statistics. The hypergeometric distribution is used for this type of sampling problem (Upton and Cook, 2014). In our context, the number of shortlisted ideas sacrificed by a random guesser follows a hypergeometric distribution. The probability of meeting the threshold by chance on a single contest is a complex function of the number of shortlisted ideas and the total number of ideas. To produce an analytical closed-form for the distribution of the count of contests for which a random guesser meets the threshold, we would require the hypergeometric distribution in conjunction with the multinomial distribution.

Rather than using the hypergeometric and multinomial distributions to produce an exact solution to the random benchmark, we use simulation for simplicity. Approximation error is negligible with a large enough number of simulation repetitions. We approximate the distribution of the number of contests for which a random guesser could meet the threshold using the following procedure:

1. For a given contest, randomly select 25% of ideas
2. Count the number of randomly selected ideas from the shortlist
3. Divide the number of selected ideas which were shortlisted by the number of shortlisted ideas. This yields the sacrifice rate.
4. Check whether the sacrifice rate is <= 15%.
5. For every contest, repeat steps 1-4 and count the number of contests which met the 15% threshold.
6. Repeat steps 1-5 10,000 times.

With 10,000 repetitions, we have an approximate distribution of results for a random guesser. We find that the median number of contests for which a random guesser meets the threshold is four contests. That means that by randomly guessing, one would tend to reach the threshold for 4 of 21 contests. The chance of a random guesser meeting the threshold for 11 of 21 contests, as the best algorithm does, is

approximately 11 in 10,000, or a chance of 0.0011. We can conclude that although the algorithms are far from perfect, they are significantly more accurate than random.

5. Discussion

This section summarizes the main results, discusses questions, draws some implications, and lists some limitations and directions for future research.

5.1 Summary of Study and Results

This paper compares three original and extended versions of algorithms from the marketing literature to evaluate ideas. We name the three algorithms: Idea's Word Colocation, Style Matching Atypicality, and Ideator's Inspiration Redundancy. Each algorithm suggests predictors of good ideas. The authors extend these algorithms and apply two methods for finding robust predictor variables: least average shrinkage and selection operator (Lasso) and K-sparse Exhaustive Search, for K<=5. The authors test the algorithms on 21 real-world crowdsourcing contests conducted for large firms. The standard provided by Hyve is "drop the worst 25% of ideas without sacrificing more than 15% of good ideas," as ranked by experts. For the analysis, we pools the 21 contests into three samples: one of large contests, one of small contests, and a third of all contests.

The main results are the following. First, acknowledging that each original algorithm provided unique theoretical and methodological value, of the three original algorithms, Inspiration Redundancy performs best out-of-sample, but fails to meet the 15% threshold. Second, for two of the three algorithms, the extended versions outperform the original. In particular, Topic Overlap Atypicality, a new measure, emerges as the most robust predictor. Third, when the best versions of the algorithms are used, all three contribute to the important out-of-sample prediction accuracy. Fourth, using extended versions of all three algorithms, we are able to meet the threshold of Hyve, to screen out the 25% of worst ideas withouot sacrificing anymore than 15% of the best ideas.

We make the following contributions to literature and practice.

First, we show the importance of testing algorithms outside of the estimation or training sample, while prior work has focused on in-sample prediction. Our results show that in-sample predictions

typically lead to better results than out-of-sample predictions across measures. Using out-of-sample prediction is a more rigorous test because in-sample predictability may exploit idiosyncrasies of the data. The finding that Inspiration Redundancy, which is an ideator-based approach, outperforms the two text-mining based approaches out-of-sample, implies that an ideator's traits are particularly important in ideation.

Second, a) We generate predictions based on a new word colocation network generated from Google Patents. The network represents the wisdom of experts and is an extension of Toubia and Netzer's (2017) wisdom of the crowd based on Google Search. b) We develop a new approach for measuring Topic Atypicality, called Topic Overlap Atypicality, whose key component is the number of topics an idea has in common with a reference corpus (divided by the number it could have in common). c) We apply Burt's (2004) Constraints measure to idea screening as an extension of Stephen, Zubcsek and Goldenberg's (2016) Inspiration Redundancy.

Third, our result that extended versions tend to outperform the original algorithms shows that one can reach important gains in the particularly important out-of-sample prediction accuracy either by developing new algorithms or by carefully expanding existing approaches. Our results are that when the best versions of the algorithms are used, all three contribute to prediction accuracy. Currently, the best algorithms in marketing can only focus on one or a few important dimensions, whereas the human mind is multi-faceted and the setting of innovation is very complex, which AI currently fails to fully capture.

Fourth, at the current stage of assessing ideas, AI cannot fully or even partially substitute for human experts. However, AI can assist humans, in our case Hyve's managers, by screening out bad ideas (Level 1), thereby reducing human's cognitive load and help them focus on the good ideas. Thus, the current study is a first steph to narrow the ideation funnel early on. Yet, as research advances and as additional algorithms bring in unique and helpful perspectives, research may develop algorithms to select the best ideas (Level 2) or even to automatically generate outstanding ideas (Level 3).

5.2 Implications

This study has important implications for managers and researchers of innovation, ideation, new products, and crowdsourcing platforms.

First, automated idea screening can be used to reduce the set of ideas evaluated by experts. This decreases experts' cognitive load and helps them make better decisions. However, this load reduction is possible mainnly by extending and combining three recently published algorithms (Berger and Packard 2018; Stephen, Zubcsek, and Goldenberg 2016; Toubia and Netzer 2017) proposed separately in the literature.

Second, as of our current state of knowledge, machine learning in real ideation contests can help in only the first performance level of AI in ideation, from least to most valuable: 1. Screening out (many) bad ideas, 2. Selection of one or a few outstanding ideas (i.e. "winners") 3. Generation of good ideas.

Third, overall, we meet the threshold provided by a knowledgeable idea contest practitioner, i.e., that not more than 15% of good ideas should be dropped to eliminate 25% of the bad ideas. This performance is good enough for Hyve to propose this screening approach to their clients. This result means that we only achieve the first step of AI in ideation, i.e., screening out bad ideas. However, although we use and extend the best methods in the marketing literature, we do not manage to achieve the second step of AI in creativity, which is to predict the winners or to automatically generate a shortlist of few ideas of which experts only have to choose the best. At the same time, we are far away from the third step of AI in creativity, which is to automatically generate good ideas; this would eliminate humans from the process of ideation. Considering that machines easily beat (almost) all humans in rule-based tasks such as chess or disease diagnosis, they are not good enough yet to accurately predict a rules-breaking task such as creativity.

Fourth, generally, we go beyond Toubia and Netzer (2017) by showing that indicators computed on the Google Patents corpus, which represents the wisdom of experts, are more often included by the lasso regression than indicators calculated on Google search, which represents the wisdom of crowds (Toubia and Netzer 2017). This finding shows that sources like Google Patents, i.e., sources which use technical language and contain expert knowledge, tend to do well in predicting expert evaluations. This

may indicate that Toubia and Netzer's (2017) approach may have broader applicability than is first apparent.

Fifth, our results lead to a re-interpretation of the contribution of Toubia and Netzer's (2017) award-winning study. Precisely, their new method and framework is generalizable across different types of networks, in our case Google Search, Google Patents, and Own networks. In our real-world dataset, for each of those three networks, the lasso regression selects different coefficients. This result suggests that the significance of coefficients likely depends on the setting and is context-specific.

Sixth, the key difference between Style Matching Atypicality (Berger and Packard 2018) and Topic Overlap Atypicality, which we develop in this paper, is that the former takes into consideration the distance between topic sets within ideas, whereas the latter does not. Therefore, the first method is sensitive to outliers, while the second is not. The fact that the Topic Overlap Atypicality is predictive while Style Matching Atypicality is not suggests that methods which are less sensitive to outliers may also be less prone to overfitting in a context of radical innovation in which uncertainty is rather high.

Seventh, while Berger and Packard (2018) found a positive relationship between atypicality and popularity, here we find a negative relationship between atypicality and the probability of being shortlisted, i.e., more typical ideas are more likely to be shortlisted. This indicates that if different ideators come up with similar ideas, the idea is likely to be promising. Kornish and Ulrich (2011) have found a similar result in a setting in which ideators cannot see each other's ideas.

5.3 Limitations

This study has limitations. First, like all such contests, the contests we study suffer from survival bias: we will never know how well the eliminated ideas would have fared had they been chosen. It is not feasible to implement thousands of ideas proposed in crowdsourcing to ascertain which does best in the market. All that we can do is to test the proposed ideas against judgements of experts. However, using such experts perhaps goes further than some prior studies, which tested their algorithms in lab studies against Mturk workers (Toubia and Netzer 2017) or peer judges in

experiments (Stephen, Zubcsek, and Goldenberg 2016). Second, we do not have extensive information about the education, interests, and work experience of the participants. Having access to such information could be used to analyze which ideators or which type of knowledge are most likely to generate great ideas.

References

Abadie, A, and Kasy, M. 2020. "The Risk of Machine Learning." *Review of Economics and Statistics*. forthcoming.

Allen, BJ, Chandrasekaran, D, and Basuroy, S. 2018. "Design Crowdsourcing: The Impact on New Product Performance of Crowdsourcing Design Solutions from the 'Crowd'." *Journal of Marketing.* 82(2): 106-123.

Amatriain, X, Lathia, N, Pujol, JM, Kwak, H, and Oliver, N. 2009. "The wisdom of the few: a collaborative filtering approach based on expert opinions from the web." *Proceedings of the 32nd International SCM SIGIR Conference*: 532-539.

Amabile. TM. 1996. *Creativity in Context: Update to the Social Psychology of Creativity.* Hachette, London, UK.

Baer, J. 1998. "The Case for Domain Specificity of Creativity." *Creativity Research Journal.* 11(2): 173-177.

Balakrishnan, V, and Lloyd-Yemoh, E. 2014. "Stemming and lemmatization: a comparison of retrieval performances." *Lecture Notes on Software Engineering.* 2(3): 262–267.

Bayus, BL. 2013. "Crowdsourcing New Product Ideas over Time: An Analysis of the Dell Ideastorm Community." *Management Science.* 59(1): 226-244.

Berger, J, Humphreys, A, Ludwig, S, Moe, WW, Netzer, O, and Schweidel, DA. 2020. "Uniting the Tribes: Using Text for Marketing Insight." *Journal of Marketing*. 84(1):1-25.

Berger, J, and Packard, G. 2018. "Are Atypical Things More Popular?" *Psychological Science.* 29(7): 1178-1184.

Bielby, WT, and Bielby, DD. 1994. "All hits are flukes: institutionalized decision making and the rhetoric of network prime-time program development." *American Journal Sociology.* 99(5): 1287-1313.

Bilalić, M, McLeod, P, and Gobet, F. 2008. "Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters." *Cognitive Psychology.* 56(2): 73-102.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, DM, Ng, AY, and Jordan, MI. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research.* 3: 993-1022.

Burroughs, JE, Moreau CP, and Mick DG. 2008. "Toward a psychology of consumer creativity." Haugtvegt, CP, Herr, PM, and Kardes, FR, eds. *Handbook of Consumer Psychology*. Lawrence Erlbaum, New York, NY. 1011-1038.

Burt, RS. 2004. "Structural Holes and Good Ideas." *American Journal of Sociology.* 110(2): 349-399.

Burt, RS. 2009. *Structural holes: The social structure of competition*. Harvard University Press, Boston, MA.

Chawla, NV, Bowyer, KW, Hall, LO, and Kegelmeyer, WP. 2002. "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research.* 16: 321-357.

Dahan, E, and Hauser, JR. 2002. "Product Development – Managing a Dispersed Process." Weitz, B, and Wensley, R, eds. *Handbook of Marketing.* Sage Publication, New York, NY, 179-222.

Dahl, DW, Chattopadhyay, A, and Gorn, GJ. 1999. "The Use of Visual Mental Imagery in New Product Design." *Journal of Marketing Research.* 36(1): 18-28.

Dahl, DW, and Moreau, CP. 2002. "The Influence and Value of Analogical Thinking During New Product Ideation." *Journal of Marketing Research.* 39(1): 47-60.

Denrell, J, and Fang, C. 2010. „Predicting the Next Big Thing: Success as a Signal of Poor Judgment." *Management Science*. 65(10): 1653-1667.

Eling, K, Griffin, A, and Langerak, F. 2014. "Using Intuition in Fuzzy Front-End Decision-Making: A Conceptual Framework." *Journal of Product Innovation Management.* 31(5): 956-972.

Friedman, J, Hastie, T, and Tibshirani, R. 2001. *The elements of statistical learning* (Vol. 1, No. 10). Springer series in statistics, New York, NY.

Gagne, C, and Shoben, EJ. 1997. "The Influence of Thematic Relations on the Comprehension of Nonpredicating Combinations." *Journal of Experimental Psychology.* 23(1): 71-87.

Giora, R. 2003. *On Our Mind: Salience, Context, and Figurative Language*. Oxford University Press, New York, NY.

Godart, FC, and Galunic, C. 2019. "Explaining the Popularity of Cultural Elements: Networks, Culture, and the Structural Embeddedness of High Fashion Trends." *Organization Science.* 30(1): 151-168.

Gobbo, C, and Chi, M. 1986. "How knowledge is structured and used by expert and novice children." *Cognitive Development.* 1(3): 221-237.

Goldenberg, J, Mazursky, S. 2002. *Creativity in Product Innovation.* Cambridge University Press, Cambridge, MA.

Greenstein, S, and Zhu, F. 2018. "Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia." *MIS Quarterly.* 42(3): 945-959.

Hanley, JA, and McNeil, BJ. 1982. „The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology. 143*(1), 29-36.

Hill, S, and Ready-Campbell, N. 2011. "The Wisdom of (Experts in) Crowds." *International Journal of Electronic Commerce.* 15(3): 73-101.

Howe, J. 2008. *Crowdsourcing. Why the power of the crowd is driving future of business.* Three Rivers Press, New York, NY.

Hughes, J, and Parkes, S. 2003. "Trends in the use of verbal protocol analysis in software engineering research." *Behavior & Information Technology.* 22(2): 127-140.

Ireland, ME, and Pennebaker, JW. 2010. "Language style matching in writing: Synchrony in essays, correspondence, and poetry." *Journal of Personality and Social Psychology. 99*(3), 549-571.

Kittur, A, and Kraut, RE. 2008. "Harnessing the wisdom of crowds in Wikipedia: Quality through coordination." *Proceedings of the 2008 ACM Conference for Computer Supported Cooperative Work.* ACM, New York, 37-46.

Kornish, LJ, and Ulrich, KT. 2011. "Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas." *Management Science.* 57(1): 107-128.

Kornish, LJ, and Ulrich, KT. 2014. "The importance of the raw idea in innovation: Testing the sow's ear hypothesis." *Journal of Marketing Research.* 51(1):14-26.

Lamm, H, and Trommsdorff, G. 2006. “Group Versus Individual Performance on Tasks Requiring Ideational Proficiency (Brainstorming): A Review.” *European Journal of Social Psychology*. 3(4): 361-388.

Luo, L, and Toubia, O. 2015. “Improving Online Idea Generation Platforms and Customizing the Task Structure on the Basis of Consumers’ Domain-Specific Knowledge.” *Journal of Marketing.* 79(5):100-114.

Matz, SC, and Netzer, O. 2017. “Big Data as a window into consumers‘ psychology.” *Current Opinion in Behavioral Sciences.* 18: 7-12.

Mobley, MI, Doares, LM, and Mumford, MD. 1992. “Process analytic models of creative capacities: Evidence for the combination and reorganization process.” *Creativity Research Journal.* 5(2): 125-155.

Netzer, O, Feldman, R, Goldenberg, J, and Fresko, M. 2012. “Mine Your Own Business: Market-Structure Surveillance Through Text Mining.” *Marketing Science.* 31(3): 521-543.

Netzer, O, Lemaire, A, and Herzenstein, M. 2019. “When Words Sweat: Written Words Can Predict Loan Default in the Text of Loan Applications.” *Journal of Marketing.* 56(6): 1-81.

Pentland, A. 2013. “Beyond the Echo Chamber.” *Harvard Business Review*. 91(11): 80-86.

Plucker, JA, and Beghetto, RA. 2004. “Why Creativity Is Domain General, Why It Looks Domain Specific, and Why the Distinction Does Not Matter.” In Sternberg, RJ, Grigorenko, EL, and Singer, JL, eds. *Creativity: From Potential to Realization*. American Psychological Association, Washington, DC, 153-167.

Stephen, A, Zubcsek, PP, and Goldenberg, J. 2016. “Lower Connectivity is better: The effects of network structure on redundancy of ideas and customer innovativeness in interdependent ideation tasks.” *Journal of Marketing Research.* 53(2): 263-279.

Stone, M. 1974. “Cross-validatory choice and assessment of statistical predictions.” *Journal of the Royal Statistical Society: Series B (Methodological).* 36(2): 111-133.

Sun, L., Zheng, X., Jin, Y., Jiang, M., & Wang, H. (2019). Estimating Promotion Effects Using Big Data: A Partially Profiled LASSO Model with Endogeneity Correction. *Decision Sciences*, *50*(4), 816-846.

Terwiesch, C, and Ulrich, KT. 2009. *Creating and Selecting Exceptional Opportunities.* Harvard University Press, Cambridge, MA.

Tibshirani, R. 1996. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society,* Series B (Methodological). 58(1): 267–288.

Toubia, O. 2006. “Idea Generation, Creativity, and Incentives.” *Marketing Science.* 25(5): 411-425.

Toubia, O, and Flores, L. 2007. “Adaptive Idea Screening Using Consumers.” *Marketing Science.* 26(3): 342-360.

Toubia, O, and Netzer, O. 2017. “Idea Generation, Creativity, and Prototypicality.” *Marketing Science.* 36(1): 1-20.

Uzzi, B, Mukherjee, S, Stringer, M, and Jones, B. 2013. “Atypical Combinations and Scientific Impact.” *Science.* 342(6157): 468-472.

Ward, TB. 1995. “What’s old about new ideas?” In Smith, SM, Ward, TB, and Fiske, RA, eds. *The Creative Cognition Approach.* MIT Press, Cambridge, MA, *157-178.*

Watts, DJ, and Strogatz, SH. 1998. “Collective dynamics of ‘small-world’ networks.” *Nature*. 393(6684): 440-442.

Wedel, M, and Kannan, PK. 2016. “Marketing Analytics for Data-Rich Environments.” *Journal of Marketing.* 80(6): 97-121.

Wessling, KS, Huber, J, and Netzer, O. 2017. “MTurk Character Misrepresentation: Assessment and Solutions. “ *Journal of Consumer Research*. 44(1): 211-230.

Wooten, JO, and Ulrich, KT. 2019. “The Impact of Visibility in Innovation Tournaments: Evidence From Field Experiments.” *Working Paper.* 1-36.

Van de Wiel, MWJ, Boshuizen, HPA, and Schmidt, HG. 2000. “Knowledge restructuring in expertise development.” *European Journal of Cognitive Psychology.* 12(3): 323-356.

Table 1. Predictor Variables.

Algorithm	Variables	Computed from which data?
TN	Mean, min, max, and coef. of variation of Jaccard indices between word pairs	Google Search, Patent Corpus
TN	Mean, min, max, and coef. of variation of node frequencies	Google Search, Patent Corpus
TN	Kolmogorov-Smirnov metric from Toubia and Netzer (2016)	Google Search, Patent Corpus
SZG	Degree	Comments network
SZG	Clustering Coefficient (a.k.a. Transitivity)	Comments network
SZG	Constraints (Burt's measure)	Comments network
BP	Topic Overlap Atypicality	All ideas from the same contest
BP	Style Matching Atypicality	All ideas from the same contest

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard

Table 2. Sacrifice Rates by Algorithm, Sample A.

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	Orig.	Ext.	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	20%	21%	23%	17%	18%	20%	18%	16%	17%
Out-of-sample	28%	23%	22%	**13%**	18%	18%	23%	17%	26%

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
Orig.: Original algorithms taken from the three publications (TN, SZG, BP)
Ext.: Extended versions of the algorithms developed in this study
Bold: meets the threshold

Table 3. Variables Used by Lasso, In-sample. DV: Shortlisted (yes/no). Sample A.

Algorithm	Variable Name	Standardized Coefficient
TN	*Node Frequency Coefficient of Variation (Patent)*	0.32
BP	*Topic Overlap Atypicality*	-0.06
TN	*Jaccard Index Coefficient of Variation (Patent)*	0.03
SZG	Degree	-0.001

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
The Lasso regression automatically excludes non-significant coefficients.
Italics: variable developed in present study

Table 4. Best variable combinations from K-sparse Exhaustive Search, Out-of-sample, Sample A.

Variable Combination	Sacrifice Rate
Topic Overlap Atypicality, Clustering Coefficient, *Avg Jaccard Index (Patent)*, *Max Jaccard Index (Patent)*, *Node Frequency Coefficient of Variation (Patent)*	9%
Topic Overlap Atypicality, Style Matching Atypicality, Clustering Coefficient, *Avg Node Frequency (Patent)*, *Node Frequency Coefficient of Variation (Patent)*	10%
Topic Overlap Atypicality, Style Matching Atypicality, Clustering Coefficient, *Node Frequency Coefficient of Variation (Patent)*, *Max Node Frequency (Patent)*	10%
Topic Overlap Atypicality, Clustering Coefficient, *Constraints Measure*, Degree, *Max Node Frequency (Patent)*	10%
Topic Overlap Atypicality, Clustering Coefficient, *Avg Node Frequency (Patent)*, *Node Frequency Coefficient of Variation (Patent)*	11%
Topic Overlap Atypicality, Clustering Coefficient, Degree, *Kolmogorov-Smirnov (Patent)*, *Avg Node Frequency (Patent)*	11%
Topic Overlap Atypicality, Clustering Coefficient, *Max Jaccard Index (Patent)*, *Node Frequency Coefficient of Variation (Patent)*	11%
Topic Overlap Atypicality, *Max Node Frequency (Patent)*	11%
Topic Overlap Atypicality, Clustering Coefficient, Degree, *Max Node Frequency (Patent)*	11%
Topic Overlap Atypicality, *Max Jaccard Index (Patent)*, *Max Node Frequency (Patent)*	11%

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
The Lasso regression automatically excludes non-significant coefficients.*Italics: variable developed or first applied in present study*

Table 5. Sacrifice Rates by Algorithm, Sample B.

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	**12%**	18%	26%	**10%**	19%	17%	**12%**	**11%**	**5%**
Out-of-sample	16%	25%	26%	**15%**	21%	**13%**	24%	16%	**15%**

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
Orig.: Original algorithms taken from the three publications (TN, SZG, BP)
Ext.: Extended versions of the algorithms developed in this study
Bold: meets the threshold

Table 6. Variables Used by Lasso, In-sample. DV: Shortlisted (yes/no). Sample B.

Algorithm	Variable Name	Standardized Coefficient
SZG	Clustering Coefficient	-0.32
BP	*Topic Overlap Atypicality*	-0.21
TN	*Kolmogorov-Smirnov Metric (Patent)*	0.20
TN	*Max Node Frequency (Patent)*	0.17
SZG	*Constraints Measure*	-0.16
TN	Jaccard Index Coefficient of Variation (Google)	0.12
TN	Node Frequency Coefficient of Variation (Google)	-0.03
TN	*Average Jaccard Index (Patent)*	0.01
TN	*Max Jaccard Index (Patent)*	-0.01

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
The Lasso regression automatically excludes non-significant coefficients.
Italics: variable developed or first applied in present study

Table 7. Best variable combinations from K-sparse Exhaustive Search, Out-of-sample, Sample B.

Variable Combination	Sacrifice Rate
Topic Overlap Atypicality, Constraints Measure, Avg Node Frequency (Patent), Min Node Frequency (Patent)	6%
Constraints Measure, Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent), Max Jaccard Index (Patent), Avg Node Frequency (Patent)	9%
Clustering Coefficient, *Constraints Measure, Avg Node Frequency (Patent)*	9%
Constraints Measure, Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent), Node Frequency Coefficient of Variation (Patent), Min Node Frequency (Patent)	9%
Constraints Measure, Degree, *Min Jaccard Index (Patent), Avg Node Frequency (Patent), Min Node Frequency (Patent)*	9%
Constraints Measure, Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent), Max Jaccard Index (Patent), Node Frequency Coefficient of Variation (Patent)	9%
Constraints Measure, Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent), Node Frequency Coefficient of Variation (Patent)	9%
Clustering Coefficient, *Constraints Measure, Kolmogorov-Smirnov (Patent), Node Frequency Coefficient of Variation (Patent)*	9%
Topic Overlap Atypicality, Constraints Measure, Jaccard Index Coefficient of Variation (Patent), Avg Node Frequency (Patent), Min Node Frequency (Patent)	10%
Topic Overlap Atypicality, Clustering Coefficient, *Constraints Measure, Avg Node Frequency (Patent), Min Node Frequency (Patent)*	10%

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
The Lasso regression automatically excludes non-significant coefficients.
Italics: variable developed or first applied in present study

Table 8. Sacrifice Rates by Algorithm, Sample C.

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	18%	18%	22%	**9%**	16%	24%	**15%**	**10%**	**9%**
Out-of-sample	27%	25%	23%	**14%**	19%	30%	24%	17%	**14%**

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
Orig.: Original algorithms taken from the three publications (TN, SZG, BP)
Ext.: Extended versions of the algorithms developed in this study
Bold: meets the threshold

Table 9. Variables Used by Lasso, In-sample. DV: Shortlisted (yes/no); Sample C.

Based on Source	Variable Name	Standardized Coefficient
BP	*Topic Overlap Atypicality*	-0.17
TN	Node Frequency Coefficient of Variation (Google)	0.02

TN = Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and Packard
The Lasso regression automatically excludes non-significant coefficients.
Italics: variable developed or first applied in present study

Table 10. Best variable combinations K-sparse Exhaustive Search, Out-of-sample, Sample C.

Variable Combination	Sacrifice Rate
Topic Overlap Atypicality, Constraints Measure, Avg Jaccard Index (Patent), Min Jaccard Index (Patent)	12%
Topic Overlap Atypicality, Constraints Measure, Min Node Frequency (Patent)	12%
Topic Overlap Atypicality, Constraints Measure, Kolmogorov-Smirnov (Patent), Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent)	12%
Topic Overlap Atypicality, Constraints Measure, Degree, *Avg Jaccard Index (Patent), Min Node Frequency (Patent)*	12%
Topic Overlap Atypicality, Constraints Measure, Min Jaccard Index (Patent)	12%
Topic Overlap Atypicality, Constraints Measure, Degree, *Min Node Frequency (Patent)*	12%
Topic Overlap Atypicality, Constraints Measure, Degree, *Avg Jaccard Index (Patent), Min Jaccard Index (Patent)*	12%
Topic Overlap Atypicality, Clustering Coefficient, *Constraints Measure,* Degree, *Min Jaccard Index (Patent)*	12%
Topic Overlap Atypicality, Constraints Measure, Degree, *Max Jaccard Index (Patent), Min Node Frequency (Patent)*	12%
Topic Overlap Atypicality, Kolmogorov-Smirnov (Patent), Jaccard Index Coefficient of Variation (Patent), Min Jaccard Index (Patent)	12%

Toubia and Netzer, SZG = Stephen, Zubcsek and Goldenberg, BP = Berger and PackardThe Lasso regression automatically excludes non-significant coefficients.
Italics: variable developed or first applied in present study

Figure 1. Lowest Quartile Sacrifice Rates, Sample A.

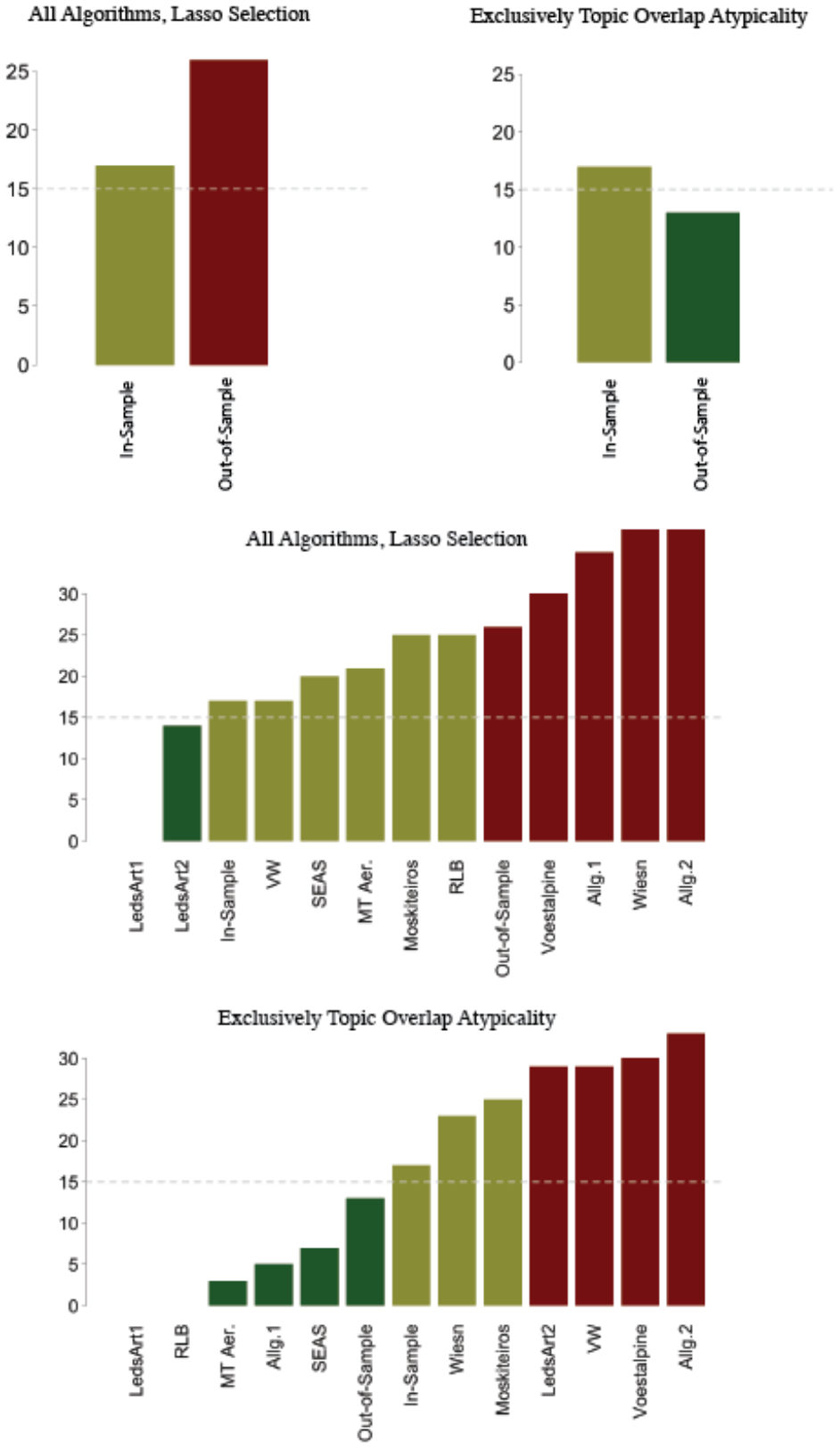


Figure 2. Lowest Quartile Sacrifice Rates, Sample B.

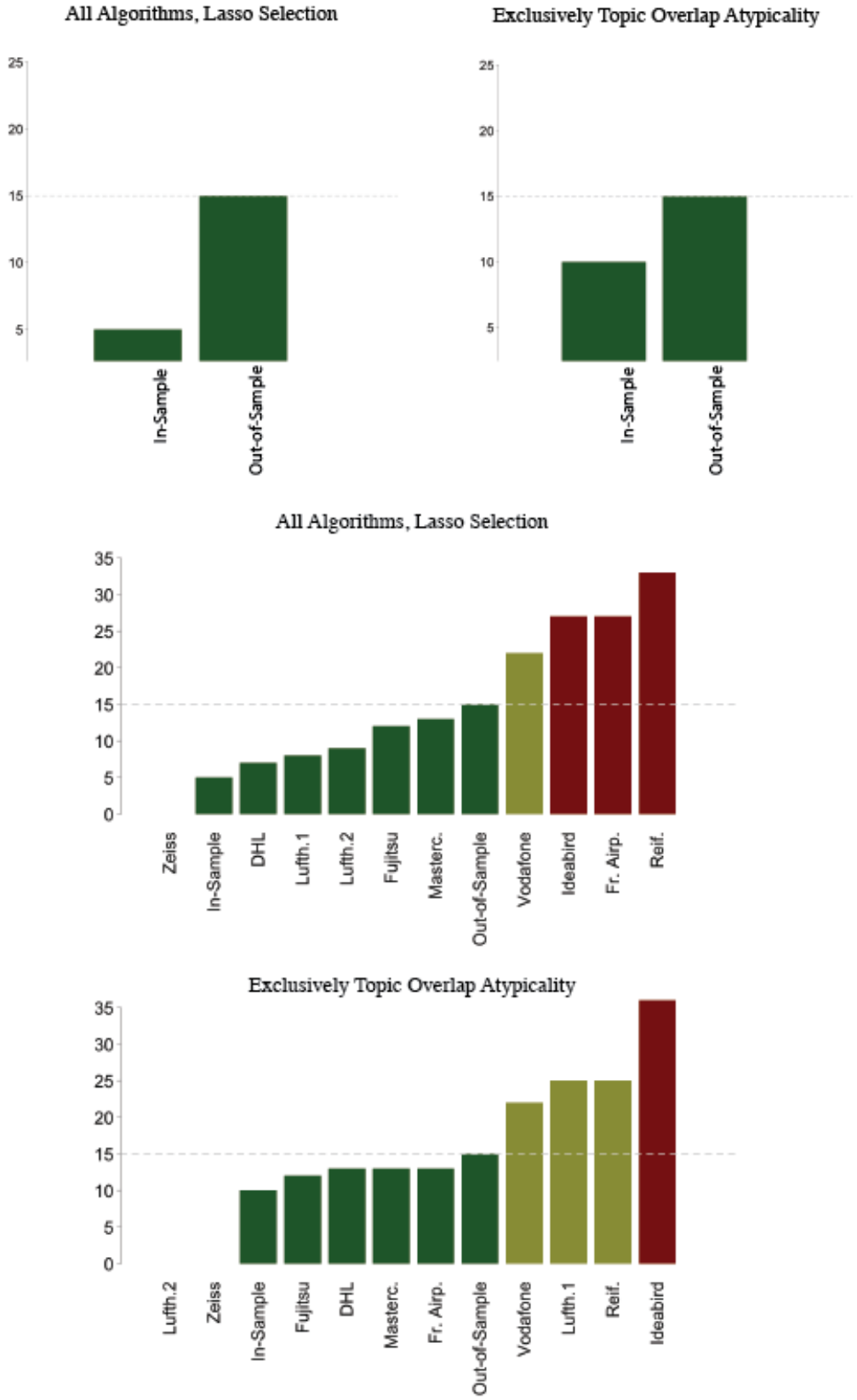
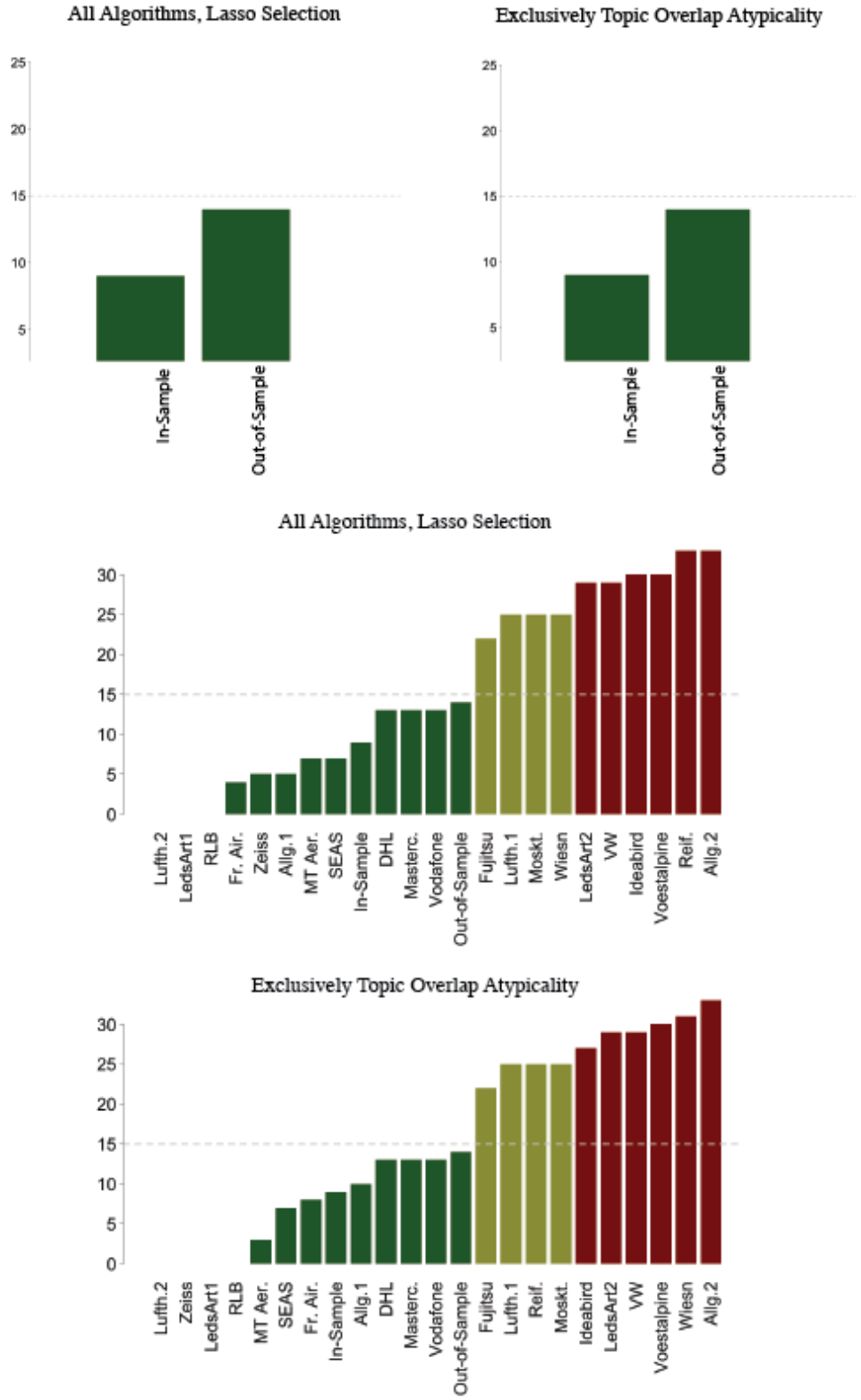


Figure 3. Lowest Quartile Sacrifice Rates, Sample C.



Web Appendix: F1 Scores

Table A1. F1 Scores by Algorithm, Sample A

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	Orig.	Ext.	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	.88	.89	.88	.88	.83	.88	.88	.89	.89
Out-of-sample	.86	.89	.86	.88	.86	.87	.86	.87	.86

Table A2. F1 Scores by Algorithm, Sample B

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	.95	.95	.95	.96	.92	.95	.95	.96	.96
Out-of-sample	.95	.95	.95	.96	.93	.94	.95	.95	.95

Table A3. F1 Scores by Algorithm, Sample C.

Test set	Word Colocation		Content Atypicality		Inspiration Redundancy		Combined		Combined Original and Extended
	Orig.	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	*Orig.*	*Ext.*	
In-sample	.94	.94	.94	.94	.81	.93	.93	.94	.94
Out-of-sample	.93	.93	.93	.94	.93	.93	.93	.94	.94