Aiding Expert Judges in Open-Innovation Challenges

YIFTACH NAGAR, Massachusetts Institute of Technology, USA
PATRICK DE BOER, University of Zurich, Switzerland
ANA CRISTINA BICHARRA GARCIA, Universidade Federal Fluminense, Brazil
KLEMENS MANG, Massachusetts Institute of Technology, USA
JAMES W. PENNEBAKER, University of Texas at Austin, USA

1. INTRODUCTION

A critical stage in open-innovation challenges is the process of screening and evaluating submissions, and selecting the best ones. It requires high-levels of expertise, especially where challenges and solutions are multi-faceted; when multiple evaluation criteria are used; and when some evaluation criteria lack objective measures. The scarcity of available expertise required for evaluation can create a bottleneck, which is exacerbated as the number of submissions grows. A notable case occurred when BP crowdsourced ideas for overcoming the DeepWater Horizon oil spill: it took a large multidisciplinary expert team several months to evaluate more than 120,000 submissions [BP, 2010]. Though this case is extreme, it exemplifies a challenge common to many open-innovation systems.

While computational evaluation will not replace expert judges in such complex scenarios anytime soon, we demonstrate how reliable predictive models can be built and employed to separate the wheat from the chaff, potentially leading to significant reduction of expert review labor. Our approach is unique in that it considers not only properties of the submission itself, but also explicit and implicit traces of human activity in relation to submissions. We developed models based on data from 18 ideation contests run within the Climate CoLab, a large-scale citizen science framework, and were able to accurately predict expert decisions about the submissions, in this set, as well as in another set of proposals submitted to 18 additional CoLab contests which were reviewed by other judges. Our approach, and many of our metrics, can be adapted to other settings.

2. METHOD

Our approach was shaped by several realizations regarding different aspects of the problem: First, while expert-panels are notoriously prone to various types of bias, due to a lack of a better alternative their judgment is still de-facto the state of the art, and widely accepted as the gold standard in studies. Second, predicting the *winners* of open-innovation challenges is hard; a more effective approach is to differentiate high quality from low quality submissions. Such a "triage" step is performed manually in many systems, and may not necessitate high-level expertise. It is probably the place where computational means can achieve the most reliability, and impact, by filtering out low-quality submissions, and freeing the experts to devote their time and skill to consider more promising submissions. Finally, previous attempts to computationally classify and rate crowd-proposals relied solely on proposals' text [e.g. Walter & Back, 2013; Westerski et al., 2013]. Yet, in open-innovation environments, traces of crowd and author activities are often available, and may provide additional clues that can help predict which proposals would be favored by expert judges.

Our resulting approach is open-ended and greedy. Based on data available in our setting, we devised a preliminary taxonomy of variables which can serve as a guideline for modeling, and which can be enhanced and appropriated to fit different settings. We developed and tested models based on this taxonomy, which take into account sociolinguistic and other aspects of proposals' text, as well as author and crowd behavior. With these models, we aim to match the reviewers' decisions at the first triage stage. We demonstrate our approach in the context of one platform – the Climate CoLab.

2.1 Setting and data

The Climate CoLab [Introne et al., 2011; www.climatecolab.org] is a sociotechnical framework designed to help people from all over the world collectively develop plans for addressing climate change. The main activity under the CoLab is a set of ideation contests that cover a wide range of technical as well as social topics related to dealing with climate change, e.g. the reduction of greenhouse gas emissions from transportation systems, geoengineering to avoid methane feedback, and urban adaptation.

Our initial dataset is comprised of the entire set of 369 proposals that were submitted to all 18 contests that ran under the Climate CoLab framework in 2012-2013. For reviewing these proposals, the CoLab organizers recruited about 60 volunteer reviewers, based on their expertise in the contest topics (about half of them were senior faculty or industry veterans, who served as *Expert Judges*; the other half were *CoLab Fellows*: graduate students and professionals, who assisted the judges). In addition, the *Crowd* (i.e. registered members of the CoLab community) could comment on proposals and indicate their support for a proposal, but had no direct role in the judgment process. The CoLab judges and fellows selected 81 proposals (~ 22%) as "Semi Finalists".

2.2 Metrics

Our taxonomy of metrics includes data of the proposal itself, and activities of authors and community members in relation to the proposal. The full taxonomy is detailed in an extended version of this paper. We grouped those metrics into the following six categories:

(1) **Readability:** Easier reading improves comprehension, retention, reading speed and readers' perseverance [DuBay, 2007]. We conjectured low readability will hinder the proposal's chance of being favored by the judges. We used several common readability measures as metrics.

(2) **Writing Style**: Social psychologists and sociolinguists have shown that people match their language, stylistically, to that of other people with whom they are communicating. Language style matching was shown to correlate with phenomena such as the strength of dyadic relationships, group cohesiveness and group task performance [Gonzales et al., 2009; Ireland & Pennebaker, 2010]. It's plausible that language style might also affect expert reviewers' perception of the proposals, and influence their decisions. The pool of reviewers of the Climate CoLab is highly educated, highly conscientious, working in academia or in knowledge work. We conjectured therefore, that as a collective, CoLab reviewers tend to have similar stylistic preferences regarding the writing of the proposals. We used LIWC [Pennebaker et al., 2007] to analyze writing style.

(3) **Potential indicators of the completeness and maturity of the proposal:** A mixed blessing in crowd-ideation contests is that with the increase in idea quantity, the share of low-quality submissions grows. We used several metrics of the text that we conjectured might provide cues regarding the completeness and maturity of the proposal: the number of hyperlinks, images, references, and whether some sections were left empty.

(4) **Length:** relates to all the former three. We counted letters, words, sentences and paragraphs.

(5) **Crowd activity:** We considered several indicators of explicit crowd activities such as the relative number of comments and "likes". In ongoing work we are also considering "honest signals" – traces of implicit crowd activities such as the number of unique page visits, time spent on a page, etc.

(6) **Author Activity:** We observed the time left from initial submission to deadline, and the number of updates in that period.

3. MODELING AND RESULTS

We first created a series of logistic-regression models for each category of predictors, using partial sets of variables that were not strongly correlated with each other. After eliminating variables that did not

have statistically-significant effects on the outcome variable, we constructed a set of integrated models, which combined the most salient predictors from all categories, and selected the final model. The final model equation is:

$$\hat{p}(semifinalist = 1|ppron, T, N, L) = \frac{1}{1 + e^{-(-2.964 - 0.379\ ppron + 1.284\ T + 0.0001\ N + 3.517\ L)}}$$

ppron is a measure of pronoun use [cf. Pennebaker et al., 2007]. Proposals with more pronouns were less likely to be selected as semi-finalists. A missing *Timeline (T)* section also lowered the chances of a proposal to be selected. Longer proposals *(N=Number of words)* had higher chances of being selected as semi-finalists by the judges, and so were proposals which received more "likes" from the crowd *(L=% of "likes" of all proposals in the contest)*. All coefficients have statistically-significant effects on the outcome ($p<0.01$).

3.1 Model evaluation and performance

To assess goodness of fit and predictive power consider the following model performance metrics: -2LL=313.04; AIC=323.05; $p>\chi^2 = 1.67E^{-15}$; Area Under the ROC Curve (AUC)=0.821.

We validated our model by building a machine-learning classifier and performing a stratified 10-fold cross-validation [cf. Kohavi, 1995]. The resulting model accuracy was 0.789 (±0.06) and the average AUC was 0.816.

We further used bootstrapping to check whether our approach can be used to build a powerful model by using only a subset of previously-judged proposals. The results, in Table 1, are encouraging:

% of data used to build the model	Resulting Area under ROC curve
80%	0.81
40%	0.80
5%	0.72

Table 1. Bootstrapping results

Finally, we ran our model on the 510 proposals sent to 18 CoLab contests the following year (2013-2014). These were mostly new contests, as were almost all the reviewers. The AUC was 0.75, re-affirming our model, and giving us further confidence that it is valid, robust, reliable, and useful. We calculated that even if we use the model in the most conservative way, maximizing sensitivity, we can skip the review of ~15% of the submissions. Allowing ~10% false-negatives (which can be handled by a secondary review process by non-experts), we can save the review of ~50% of the submissions.

4. CONCLUSION

The bottleneck of expertise for reviewing a mass of complex ideas submitted to open-innovation platforms poses a barrier to innovation. Here we demonstrate, on a live, large-scale system, with tens of thousands of members, and hundreds of crowd submissions, that it is possible to use a limited subset of manually judged submissions, to build and train a classifier that can reliably aid in reducing the load on expert judges. We continue working on modeling proposal success in the Climate CoLab, observing additional years, and enhancing our taxonomy, our modeling, and our process. Further modeling work in different settings will help strengthen the external validity of our results, and provide insight regarding which variables, or families of variables, are important to look at in any settings, and which are context specific. We will gladly offer our taxonomy, advice, and assistance in implementing our approach, to others who may be interested in joining us in pushing open innovation forward.

REFERENCES

BP. (2010). Deepwater Horizon Containment and Response: Harnessing Capabilities and Lessons Learned. .

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*: ERIC.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*.

Introne, J., Laubacher, R. J., Olson, G. M., & Malone, T. W. (2011). *The Climate CoLab: Large scale model-based collaborative planning*. In proceedings of the Conference on Collaboration Technologies and Systems (CST 2011), Philadelphia, PA.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology, 99*(3), 549-571. doi: 10.1037/a0020386

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* In proceedings of the International Joint Conference on Artificial Intelligence.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC. Net.

Walter, T. P., & Back, A. (2013). *A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests.* In proceedings of the 46th Hawaii International Conference onSystem Sciences (HICSS).

Westerski, A., Dalamagas, T., & Iglesias, C. A. (2013). Classifying and comparing community innovation in Idea Management Systems. *Decision Support Systems, 54*(3), 1316-1326. doi: http://dx.doi.org/10.1016/j.dss.2012.12.004