Content-based Success Prediction of Crowdfunding Campaigns: A Deep Learning Approach

SeungHun Lee
Sangmyung University
Cheonan, South Korea
mr.leesh90@gmail.com

KangHee Lee
Sangmyung University
Cheonan, South Korea
lkh7054@gmail.com

Hyun-chul Kim*
Sangmyung University
Cheonan, South Korea
hyunchulk@gmail.com

Abstract
Despite the huge success of crowdfunding platforms, the average project success rate is 41%, and it has been decreasing. Hence, finding out the factors that lead to successful fundraising and predicting the probability of success for a project has been one of the most important challenges in the crowdfunding. This work is the first attempt to use in-band project content - text - data only, contained in all the *Campaign*, *Updates*, and *Comments* sections of a crowdfunding project (not in combination with any other out-of-band project metadata or statistically-derived numeric features), for success prediction. By adopting (i) the sequence to sequence (seq2seq) deep neural network model with sentence-level attention and (ii) Hierarchical Attention-based Network (HAN) model, we demonstrate that our proposed model achieves the state-of-the-art performance in predicting success of campaigns, as much as 89-91%. We also show that our method achieves 76% accuracy on average on the very first day of project launch, using campaign main text data only.

Author Keywords
Crowdfunding; Kickstarter; Deep Learning; Success Prediction; Natural Language Processing.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g.,HCI)]: Miscellaneous

Introduction
Crowdfunding, a newly emerging practice of online fundraising from people by advertising project ideas, has emerged as a lucrative alternative in acquiring investments for new startups facing the daunting challenges of financing. As an alternative to traditional venture capital investment, crowdfunding manages to become an unwavering support for the individuals, small businesses, startups, and industries

*Corresponding Author: Hyun-chul Kim (hyunchulk@gmail.com)

CSCW '18 Companion, November 3–7, 2018, Jersey City, NJ, USA
ACM 978-1-4503-6018-0/18/11.
https://doi.org/10.1145/3272973.3274053

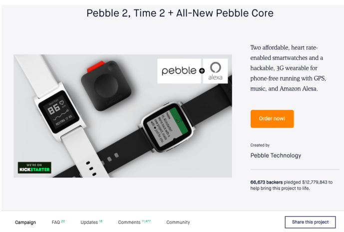

Figure 1: A screenshot of an example Kickstarter project: Pebble smartwatch

Section	The average # of sentences
Campaign	40.7
Updates	92.4
Backers comments	375.5
Creator comments	91.85
Speech	1.9

Table 1: The average number of sentences per project in technology

by soliciting huge amounts from a large number of people. However, even though the number of projects and amount of pledged money on crowdfunding platforms has tremendously grown, the average project success rate, i.e., the percentage of projects that successfully achieve their desired funding goal, is 41% (Table 2), and it has been decreasing [2]. Hence, finding out the factors that lead to successful fundraising and predicting the probability of success for a project has been one of the most important challenges in the crowdfunding. In recent years, several studies have investigated what factors lead to a fundraising success or failure, and how influential the factors are [6, 3]. Mollick found that generic static features of projects (e.g., the presence or absence of a campaign video, spelling errors, frequent updates, etc.) and social network-related features like the number of Facebook friends of founders are strongly related to the success of a project [3]. Xu et al. found that specific uses like types, themes and timing of updates have powerful associations with project success [6]. It also has been reported that text phrases in *Update* [6] as well as static or statistically-derived numeric features such as the number of comments [4], images, videos, users, backers, etc., [2] are key success-determinants of crowdfunding projects.

However, prior work has mostly focused on using (i) out-of-band project metadata such as project category, duration, goal money, geographic location, etc., or (ii) statistically-derived numeric features capturing behavioral, social networking, update patterns or linguistic cues of backers and creators; such as the number of backers, updates, comments, sentences, words, Facebook friends, etc. To the best of our knowledge, this work is the first attempt to use in-band project content - text - data themselves only, for success prediction of crowdfunding projects. By adopting (i) the sequence to sequence (seq2seq) deep neural network model [5] with sentence-level attention and (ii) Hierarchical Attention-based Network (HAN) model [7], we show our proposed model achieves the state-of-the-art performance in predicting success of campaigns, as much as 89-91%.

Methodology

We collected all the pages of 216,136 crowdfunding campaigns available on Kickstarter.com, launched between April 2009 and August 2017. Out of all the 15 main categories predefined by Kickstarter as shown in Table 2, we chose the Technology category for our analysis, as (i) it is one of the top 3 active categories on Kickstarter, representing around 24% of all the money pledged in 2017 [1], and (ii) according to our analysis (in Table 2), projects in the Technology category are under more intense competition; only 20.4% of projects in the category meets its funding goals, which we believe makes a project success prediction service more attractive, once made. From that category, we randomly sampled 1,000 successfully-funded projects and 1,000 failed projects for our experiments.

Figure 1 shows an example Kickstarter campaign, which has 1) *Campaign* section where the project creator introduces and describes the project idea often with the help of images and videos, 2) *Updates* section where project creator keeps their backers (i.e., supporters) updated with the project progress, 3) *Comments* section where creator communicate with backers. For each campaign, we collected all the text content available in the *Campaign*, *Updates* and *Comments*. We also collected 4) *Speech* content (spoken words) extracted from campaign videos posted, using automatic speech recognition service provided by Microsoft[2].

[1] http://icopartners.com/2018/01/kickstarter-2017-year-review/
[2] https://azure.microsoft.com/ko-kr/services/cognitive-services/

Category	Total # of projects	successful projects(%)
Journalism	2,752	18.1
Technology	22,982	20.4
Food	11,978	21.4
Crafts	3,714	22.6
Fashion	11,938	23.8
Photography	4,624	28.1
Publishing	29,945	32.7
Art	18,622	40.7
Games	19,180	50.1
Music	34,190	50.2
Dance	1,304	55.2
Film & video	35,781	55.4
Theater	3,538	55.7
Comics	5,021	56.8
Design	10,567	76.7
Total	216,136	41.8

Table 2: The total number of projects and percentage of successfully funded projects in our dataset (2009-2017)

We attempt to solve the problem of campaign success prediction by adopting (i) the seq2seq neural network model [5] with sentence-level attention and (ii) HAN model [7]. Both methods have achieved great success in a variety of NLP tasks like machine translation and speech recognition,as training models for converting sequences from one domain (e.g., sentences in French) to sequences in another (e.g., the same sentences translated in English). Using both methods, our text datasets collected from *Campaign*, *Updates*, *Comments*, and *Speech* are mapped into a vector representation. We implemented them with Gated Recurrent Units (GRU) [1]. For word embedding, we use pre-trained 300-dimensional GloVe [3]. Our models are trained for 7 epochs on batch size of 32 with the RMSprop optimizer and learning rate of 0.001. We randomly split 0.8, 0.1 and 0.1 of dataset for training, validation, and testing.

Preliminary Results

The results are shown in Table 3. It turns out that text data written in *Campaign*, *Updates*, and *Comments* sections contain valuable information in predicting chances of successful fundraising. Our models built with *Updates* and backers' *Comments* achieve 85-91% of accuracy, the state-of-the-art performance in predicting success of campaigns, particularly with only text data from each section. To the best of our knowledge, this is the first work that empirically discovers how helpful the contents of backers' comments (i.e. text) themselves are, in predicting project success. It has been already known that phrases in *Update* [6] and the number of comments [4] are key success-determinants of crowdfunding projects. The *Speech* data were not helping any in prediction, with 49.5-50% accuracy; on average, only 1.9 sentences were extracted from campaign title videos, which apparently were not large enough to build an effective predictive model (See Table 1).

[3] https://nlp.stanford.edu/projects/glove/

Section	Seq2seq [5] + Attention	HAN [7]
Campaign	76	62.5
Updates	91	85
Backers *comments*	86.5	89
Creator *comments*	80.5	79.5
Speech	49.5	50
Backers + Creator *comments*	88.5	82.5
Campaign + *Speech*	71.5	58
Campaign + *Updates* + All *comments* + *Speech*	80	60

Table 3: Performance camparisons of classification of two methods

We further explore how early our seq2seq+attention models can predict success of a campaign as it progresses. Figure 2 plots the average prediction accuracy over elapsed time after project launch. On the very first day (i.e., 0-day) of the project start, with only *Campaign* section text available, our model achieves 76% of accuracy. This result is comparable to or higher than those of previous work like [2], where 0-day predictions were made with *Campaign* section data, based on static or statistical features such as the number of images, sentences, videos, etc. While *Campaign*-based model performs better in the earliest phase of day 0-5, both *Comments*-based and *Update*-based predictors start to outperform it after 5-10 days and 20-30 days, respectively, as more and more comments and updates arrive as time goes by. After a month of project launch, *Update* and backers' *Comments* stay as the best predictors with constantly and slightly increasing accuracies, until the end of a project duration.

Our ongoing work includes: (i) we need to rationalize the outcomes from the built models, using various techniques like First-derivative Saliency and Feature Erasing or Selective Adding, which have been widely used for visualizing and understanding neural models, (ii) in-depth investigation

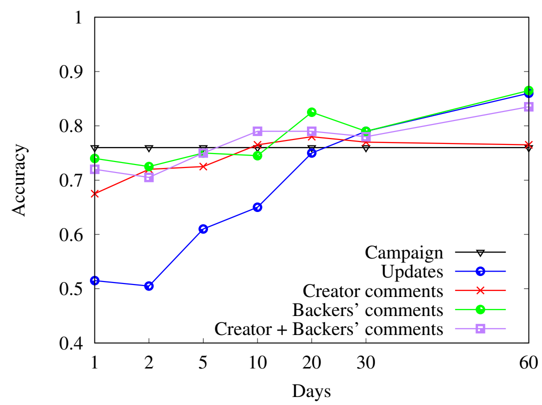

Figure 2: Estimated average prediction accuracy (seq2seq + attention) vs. elapsed time(days)

into potential synergy between our content-based methods and the other conventional (numeric) feature-based approaches (e.g., the number of videos, comments, updates, etc.), and (iii) applying the proposed methods to the problem of detecting deceptive or fraudulent campaigns as well.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016R1D1A1A09919378).

REFERENCES

1. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
2. Jinwook Chung and Kyumin Lee. 2015. A long-term study of a crowdfunding platform: Predicting project success and fundraising amount. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 211–220.
3. Ethan Mollick. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of business venturing* 29, 1 (2014), 1–16.
4. Thomas Müllerleile and Dieter William Joenssen. 2015. Key success-determinants of crowdfunded projects: An exploratory analysis. In *Data science, learning by latent structures, and knowledge discovery*. Springer, 271–281.
5. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
6. Anbang Xu, Xiao Yang, Huaming Rao, Wai-Tat Fu, Shih-Wen Huang, and Brian P Bailey. 2014. Show me the money!: An analysis of project updates during crowdfunding campaigns. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 591–600.
7. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.