2019 IEEE 6th International Conference on Industrial Engineering and Applications

Finding the Keywords Affecting the Success of Crowdfunding Projects

Long-Sheng Chen
Department of Information Management,
Chaoyang University of Technology,
Taichung, Taiwan
e-mail: lschen@cyut.edu.tw

En-Li Shen
Department of Information Management,
Chaoyang University of Technology,
Taichung, Taiwan
e-mail: s10714616@gm.cyut.edu.tw

***Abstract*—Because the rise of crowdfunding, entrepreneurs decrease seeking help from traditionally financial institutions, but began to get help on the Internet. Now, more than 15 million people involved and the amount of funds raised exceeded 3.9 billion US dollars. Although crowdfunding provides a new fundraising channel for entrepreneurs who need to raise funds, success in reaching the target amount is a big challenge. How to increase the success rate of fundraising projects is the most concern of all fundraisers. Most of the current researches aimed to explore the relation between the founders and the success of the project. Relatively few works focus on the impact of the description and the wording of the project to predict the success rate of fundraising. Therefore, this study will collect real fundraising projects from Kickstarter, and analyze the text description content of these projects. The feature selection method, Support Vector Machines Recursive Feature Elimination (SVM-RFE), has been employed to find key words that may affect the success of the project. Then, we'll use selected keywords to build a prediction model by utilizing Support Vector Machines (SVM) to help emerging entrepreneurs or anyone who needs to raise funds can have a higher chance of successful fundraising.**

Keywords-crowdfunding; feature selection; project success rate; prediction; SVM-RFE

I. INTRODUCTION

Before the year 2000, entrepreneurs often encountered insufficient resources in the early stage of their business, including insufficient funds, investor information asymmetry and lack of guarantors [1]. However, because of the rise of crowdfunding, entrepreneurs decrease seeking help from financial institutions such as investment angels or banks, but began to use the power of the masses on the Internet to get help [2].

According to the report of Kickstarter which is the world's largest public fundraising platform, there were 420,000 fundraising projects by October 2018, including 150,000 successful projects that reached the target amount [3]. Totally more than 15 million people involved and the amount of funds raised exceeded 3.9 billion US dollars. It can be seen that in this era of social networks, crowdfunding has become one of the most important fundraising methods.

Although crowdfunding provides a new fundraising channel for entrepreneurs or those who need to raise funds, success in reaching the target amount is another challenge. How to increase the success rate of fundraising projects is the most concern of all fundraisers. Lots of published literatures indicated that some factors will affect the success of fundraising. For example, some studies indicated the identity of the founder of the project, the experience of the founder of the project, the number of comments on the project, the number of updates to the project, and the description of the project, will affect the success rate of fundraising projects [4]. The level of sophistication, fundraising time and funding goals are closely related to the success of the crowdfunding project, and the project introduction film also has an important position in the entire project [5]. Other study also proposed a framework that includes the use of five project attributes for analyzing and predicting success rate [6].

To sum up, most of the current researches aimed to explore the characteristics of the founders who influence the success of the project and the characteristics of the project. Relatively few works focus on the impact of the description and the wording of the project to predict the success rate of fundraising. Therefore, this study will collect real fundraising projects from Kickstarter, and analyze the text description content of these projects. The feature selection method, Support Vector Machines Recursive Feature Elimination (SVM-RFE), has been employed to find key words that may affect the success of the project. Then, we'll use selected keywords to build a prediction model by utilizing Support Vector Machines (SVM) to help emerging entrepreneurs or anyone who needs to raise funds can have a higher chance of successful fundraising.

II. LITERATURE REVIEW

A. Crowdfunding

Crowdfunding is a different approach to traditional fundraising. It allows individuals, groups or companies to build innovative ideas, products or services through the power of the crowd based on fundraising sites. Crowdfunding can be regarded as results of the generalization and electronic way of traditional fundraising. It can not only replace the funds and help from a few people, but can also raise ideas, products or services to the public through fundraising on the Internet, to achieve the marketing effect. The process and results of implementing crowdfunding can be regarded as indicators of new products or services before they enter the market [7].

The current fundraising platform can be roughly divided into four types: debt, equity, donation and reward. Debt is essentially a loan relationship. Investors lend money to the

978-1-7281-0851-3/19/$31.00 ©2019 IEEE

creator, while creators need to give interest and return principal according to the agreement. Equity refers to the creator's need to propose part of the equity as the cost of raising funds. Donation is that donors do not ask for returns. It's a public welfare contribution. Reward is the most common type. Investors invest in creators to obtain future or exclusive limited products and services [8-9].

Depending on the fundraising model, the crowdfunding platform also can be roughly divided into two categorizations, All-Or-Nothing and Keep-It-All. In the former one, the fundraiser can get funds within the time limit, if the target amount is achieved. If the target is not achieved, no funds will be obtained. The latter will be able to obtain the funds even if the fundraiser does not reach the target amount. Most of the world's crowdfunding platforms in the world now use "All-Or-Nothing" [10].

B. Text Mining

Text mining aims to find key and useful information from the document, and to further analyze the information. Due to the rapid increase in the amount of information on the Internet, these unstructured or semi-structured texts need to be processed using text mining techniques to identify the structures and rules that are hidden in them [11].

Some studies attempt to use text mining to find out the potential characteristics of the text data, such as analysis of the content of the description, the influence of the project founder's emotions in the production of the project description on the successful crowdfunding project [12], the analysis of the quality of the argument and the source credibility of the project description on the success of the crowdfunding project [13].

In addition, previous works often use questionnaires to survey data. But, traditional questionnaires have experimental effects and sampling bias. The information in online texts is more objective, massive, and has no sample bias [14]. Therefore, this study will use text mining to analyze data.

C. Feature Selection

Feature selection is a commonly used technique in data mining. In addition to significantly reducing feature space, it can also improve the predictive accuracy of classifiers by eliminating unimportant and irrelevant features to obtain better classification results [15]. SVM-RFE is one of feature selection techniques in data mining, and can find important information from a large number of features. It arranges all features in descending order through the obtained weight vector w. The more important the ranking is, the more important it is to indicate this feature [16]. SVM-RFE is widely used, such as cancer prediction in medicine [17], protein subcellular localization for biomedical applications [18], and prediction of electricity prices in electricity market analysis [19]. Therefore, this study will use SVM-RFE as a feature selection method to explore the ranking of important words in the introduction of crowdfunding projects.

D. Support Vector Machines

Cortes and Vapnik [20] proposed Support Vector Machines (SVM) in 1995, mainly for supervised learning models and related learning algorithms for analyzing data in classification and regression analysis. Original SVM is a binary classifier, which is mainly used to deal with the problem of two or more classes of data classification, and uses the hyperplane of the largest boundary as the decision surface [21]. Lots of literatures reported that SVM has a good performance in identifying sentiment, including identifying electronic product reviews [22], and deceptive spam [23-24]. Therefore, this study will use SVM as a classification method to confirm the result of feature selection and build a prediction model.

III. METHODOLOGY

The experimental process of this study is shown in Figure 1. It is divided into six steps.

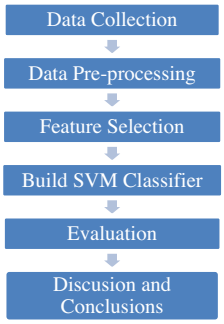

Figure 1. The implemental process of this study

A. Data Collection

This study collects data from Kickstarter (https://www.kickstarter.com/). We collect "game" related project descriptions, and use the website crawler tool to extract texts of the project description.

B. Data Pre-Processing

First, all words which contain special characters and non-English will be removed to avoid errors in the pre-experimental processing. Then, before counting the word frequency, in order to avoid collecting a large number of redundant words, this study also screen stop words, such as "a", "the", "and", "or", "with", etc. Next, we kept words whose frequency is above 30, and the Term-Document Matrix (TDM) is established by using TF-IDF weights. Before implementing feature selection, all the data is normalized to the interval [0, 1]. Finally, we give labels to pre-processed data. Y=1 means that the project achievement rate is 100% or more, and Y=0 means that the project achievement rate is 75% or less.

C. Feature Selection

Firstly, the five-fold cross-validation experiment has been used. The experimental data is divided into five equal parts, four of which are training subsets and the other one is a test subset in turn. Therefore, after implementing feature

selection, we will have five feature subsets, and we will select optimal subset depending on occurrence frequency.

This study will use the support vector machines recursive feature elimination (SVM-RFE) to select important features. This method is based on the SVM classifier training algorithm. The obtained weight vector w uses the training ordering coefficient. The minimum weight of the feature vector is removed, and the new feature combination of the remaining training sets will be reclassified by repeating the above steps. The complete steps are as follows:

Step 1 First, the training sample matrix is performed: $X_0 = [X_1, X_2, \cdots X_K, \cdots, X_i]^T$, category label: $y = [Y_1, Y_2, \cdots Y_K, \cdots, Y_i]^T$.

Step 2 Initialize the original feature set $s = [1,2, \cdots, n]$.

Step 3 lists feature set r according to feature criteria $r = [\,]$.

Step 4 Next, repeat the following steps from 4.1 to 4.5 until $s = [\,]$.

Step 4.1 Split the training sample into features $X_0 = X(:, s)$

Step 4.2 Given the parameter values and then training the classifier $\alpha = SVMtrain(X_0, \mathrm{Y}, \mathrm{C}, \gamma)$, where C is the penalty factor and γ is the core function of RBF.

Step 4.3 Calculate the ranking coefficient $C_{\mathrm{i}} = W_i 2$; if (s. length > 1000) Quadratic partitioning of the dynamic limit, else Delete the minimum characteristic of the ranking coefficient.

Step 4.4 Update Feature List $r = [s(f), r]$, s(f) is the feature set for preliminary deletion in the iteration.

Step 4.5 Exclude the characteristics of the minimum coefficient, and establish the residual feature set after exclusion, the formula is as follows:

$$s = s(1: f[0] - 1, f[lenghth(f)]: length(s))$$

Step 5 Output the sorted feature list: r.

D. Build SVM Classifier

The selected features in last step will be evaluated by SVM. And we also can build a SVM prediction model using the best feature subsets. The steps of implementing SVM are as follows.

Step 1: Normalize Raw Data

Step 2: Transform the data into a format that conforms to the SVM tool

Step 3: Use the RBF kernel function, as in Equation (1).

$$\mathrm{K}(\mathrm{x}, \mathrm{y}) = \mathrm{e}^{-\gamma||\mathrm{x}-\mathrm{y}||^2} \tag{1}$$

Step 4: Use cross-validation to select the optimization parameters C and γ.

Step 5: The obtained optimization parameters C and γ are used to train and obtain the SVM model.

Step 6: Put the test data into the trained SVM model for testing.

E. Evaluation

We will use the evaluation indicators including Positive Accuracy (PA), Negative Accuracy (NA), G-Mean (GM), Overall Accuracy (OA), and F1-Measure (F1). Finally, the best performance results are selected to find the best classification subset.

The training results will be calculated and evaluated by the Confusion Matrix. As shown in Table I, TP and TN are the correct number of samples for positive and negative samples, respectively. FP and FN are negative and positive examples, but misclassified into positive and negative, respectively.

TABLE I. CONFUSION MATRIX

Predicted / Actual	Positive	Negative
Positive	True Positive, TP	False Negative, FN
Negative	False Positive, FP	True Negative, TN

The calculation of the Positive Accuracy (PA) and the Negative Accuracy (NA) are as shown in Equation (2)~(3).

$$\mathrm{PA} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{NA} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TN}} \tag{3}$$

Finally, PA and NA are used to calculate geometric mean (G-Mean, GM), overall accuracy (OA) and F1 measure (F1) as important indicators for classification performance evaluation, such as equations (4)~(6).

$$\mathrm{GM} = \sqrt{\mathrm{PA} \times \mathrm{NA}} \tag{4}$$

$$OA = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{5}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

F. Discussion and Conclusions

In this study, all feature subsets and SVM classification results of the original feature set are evaluated and analyzed, and the best classification subset is found according to the classification results. Based on the experimental results, this study will provide some suggestions to the owners of crowdfunding projects as reference for increasing success rate in the future.

IV. EXPERIMENTAL RESULTS

A. Data Collection and Pre-Processing

This study uses the "Python" language to program a crawler tool to extract the text data of the crowdfunding project description, from the famous public fundraising website Kickstarter (https://www.kickstarter.com/) shown in Figure 2. We focus on "Game" categorization projects, and collect projects description as Figure 3 from July 1, 2018 to August 30, 2018. We filter the projects by their success rate. We collected 50 projects which achieve above 100% target

amount, and another 50 projects which achievement rate is under 75%.

This study first removes the special characters and non-English words from the crawled data to avoid errors in the subsequent pre-processing. Then, we use the data mining tool "QDA Miner" to perform word frequency statistics, and retain the words whose frequency is more than 30 times.

Figure 2. Kickstarter Homepage

Figure 3. An example of project introduction

B. Experimental Results

This study uses software tools for learning. "Weka 3.8" and "libSVM" have been used to implement SVM-RFE and build SVM model, respectively. In SVM-RFE, we can get ranking of features. We merely extract top 20% features. According to the occurrence frequency of features, we obtain 2 feature subsets. They are SVM-RFE # 1 (the number of occurrences in the top 50 is more than 5 times) and SVM-RFE # 2 (the number of occurrences in the top 50 is more than 3 times). The amount of features in SVM-RFE # 1 and SVM-RFE # 2 is 45 words and 50 words, individually.

Tables II & III list the evaluation results of feature subsets, SVM-RFE #1 & #2. Table IV provides the comparison between original feature set and other 2 selected feature sets.

It can be seen from Table IV that SVM-RFE #2 has the best performance, and all indicators are superior to SVM-RFE #1 feature subset and original feature sets. Therefore, this study uses SVM-RFE #2 feature set as the best classification subset.

TABLE II. SVM-RFE FEATURE SET #1

Index	Fold1	Fold2	Fold3	Fold4	Fold5	Mean	StDev
PA	88.9%	85.7%	57.1%	100%	66.7%	79.7%	17.4%
NA	81.8%	69.2%	53.9%	62.5%	63.6%	66.2%	10.3%
GM	85.0%	75.0%	55.9%	70.0%	65.0%	70.0%	11.2%
OA	85.3%	77.0%	55.5%	79.1%	65.1%	72.4%	11.9%
F1	84.2%	70.6%	47.1%	57.1%	63.2%	64.4%	14.0%
Time (sec)	0.01	0.01	0.04	0.01	0.01	0.02	0.01

TABLE III. SVM-RFE FEATURE SET #2

Index	Fold1	Fold2	Fold3	Fold4	Fold5	Mean	StDev
PA	100%	100%	85.7%	100%	60.0%	89.1%	17.4%
NA	83.3%	71.4%	69.2%	66.7%	60.0%	70.1%	8.5%
GM	90.0%	80.0%	75.0%	75.0%	60.0%	76.0%	10.8%
OA	91.3%	84.5%	77.0%	81.7%	60.0%	78.9%	11.8%
F1	88.9%	75.0%	70.6%	66.7%	60.0%	72.2%	10.8%
Time (sec)	0.01	0.01	0.02	0.02	0.02	0.02	0.01

TABLE IV. COMPARISON OF SELECTED FEATURE SUBSETS EVALUATED BY SVM

Feature set	Original (230)	SVM-RFE #1 (45)	SVM-RFE #2 (50)
Index	Mean (St.Dev.)	Mean (St.Dev.)	Mean (St.Dev.)
PA(%)	74.00% (13.23%)	79.68% (17.40%)	89.14(17.43)
NA(%)	66.88% (14.69%)	66.21% (10.32%)	70.13(8.54)
GM(%)	69.00% (13.87%)	70.00% (11.18%)	76.00(10.84)
OA(%)	70.25% (13.41%)	72.39% (11.95%)	78.90(11.76)
F1(%)	63.93% (17.38%)	64.43% (14.01%)	72.23(10.82)
Time (sec.)	0.03 (0.02)	0.02 (0.01)	0.02 (0.01)

According to the results, 50 words that affect the success of the fundraising project were found, as shown in Table V. From Table V, we merely focus on top 25 keywords, and still have some findings. We can found that the keywords that affect the project success could be divided into several groups. They are 'guarantee' (pledge, always, risks, check), 'rewards' (rewards, included, shipping), 'game design and content' (run, stretch, hand, rules, cards), and 'products' (product, quality, price).

V. CONCLUSIONS AND FUTURE WORKS

The purpose of this study is to discover the keywords that affect the success of the project from the introduction description of the fundraising projects. SVM-RFE has been used to screen out the important features, then the SVM is used to build prediction model. From results of SVM-RFE feature selection, we can indicate 50 important keywords. And they can be grouped as 'guarantee' (pledge, always, risks, check), 'rewards' (rewards, included, shipping), 'game design and content' (run, stretch, hand, rules, cards), and 'products' (product, quality, price). Based on the selected important keywords, we also build a SVM prediction model. Averagely, this model can have 78.90% accuracy for predicting the success of crowdfunding projects. According to the above conclusions, we suggest that when game producers or manufacturers create a crowdfuning project, the

project description should be focused on ‘game design and content’, ‘products’, ‘guarantees’, and ‘reward’.

TABLE V. SELECTED KEYWORDS OF AFFECTING SUCCESS RATE BY SVM-RFE

1	RUN	18	INCLUDE	35	BASE
2	KICKSTARTER	19	PLEASE	36	BOOK
3	PLEDGE	20	RULES	37	FINAL
4	ALWAYS	21	CHECK	38	ART
5	STRETCH	22	BOX	39	GOALS
6	HAND	23	ADDITIONAL	40	KEEP
7	REWARDS	24	LITTLE	41	PROJECTS
8	CAMPAIGN	25	CARDS	42	SHIP
9	LOOK	26	POSSIBLE	43	NUMBER
10	BACKERS	27	ADD	44	CARD
11	SHIPPING	28	PRICE	45	DELIVERY
12	INCLUDED	29	COMES	46	PRINT
13	INCLUDING	30	RECEIVE	47	TOTAL
14	RISKS	31	PAGE	48	FULL
15	PRINTING	32	COPY	49	ADVENTURE
16	PRODUCT	33	PRODUCTS	50	FAMILY
17	QUALITY	34	COPIES		

For the potential directions of future works, we have some suggestions. First, this study only uses the Kickstarter platform as our data source. In the future, it is necessary to collect more data from different crowdfunding platforms in different countries to enhance the effectiveness of the experiment. Second, the total experimental data of this study is 100, including 50 successful projects and 50 failed projects. The data size could be improved. Finally, this study only uses SVM-RFE for feature selection and SVM classifier for learning. In the future, more different feature selection methods and classification methods can be added.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Council of Taiwan, R.O.C. (Grant No. MOST 107-2410-H-324 -004).

REFERENCES

[1] A. Cosh, D. Cumming, & A. Hughes, “Outside entrepreneurial capital,” The Economic Journal, vol. 119,no. 540, 2009, pp. 1494-1533.

[2] F. Kleemann, G. G. VoB, & K. Rieder, “Un(der) paid innovators: The commercial utilization of consumer work through crowdsourcing,” Science, Technology & Innovation Studies, vol. 4, no. 1, 2008, pp. 5-26.

[3] Kickstarter, https://www.kickstarter.com.

[4] T. Kim, M. H. Por, & S.-B. Yang, “Winning the crowd in online fundraising platforms: The roles of founder and project features,” Electronic Commerce Research and Applications ,Vol. 25, 2017, pp. 86-94.

[5] R. Fernandes, “Analysis of crowdfunding descriptions for technology projects,” Doctoral dissertation, Massachusetts Institute of Technology, 2013.

[6] H. Ynan, R. Y. K. Lau, &W. Xu, “The determinants of crowdfunding success: A semantic text analytics approach,” Decision Support Systems, vol. 91, 2016, pp. 67-76.

[7] E. Mollick, “The dynamics of crowdfunding: An exploratory study”, Journal of Business Venturing, vol. 29, no. 1, 2014, pp. 1-16.

[8] J. Hollas, “Is Crowdfunding now a threat to traditional finance? ,” Corporate Finance Review, vol. 18, no. 1, 2013, pp. 27-31.

[9] D. Colgren, 2014, “The rise of crowdfunding: Social media, big data, cloud technologies,” Strategic Finance, vol. 96, no. 4, 2014, pp. 56-57.

[10] J. Hui, E. Gerber, & M. Greenberg, “Easy Money? The Demands of Crowdfunding Work,” Technical Report No. 4, Segal Design Institute, Northwestern University, 2012.

[11] G. M. Thomaz, A. A. Biz, E. M. Bettoni, L. Mendes-Filho, & D. Buhalis, “Content mining framework in social media: A FIFA world cup 2014 case analysis,” Information and Management, vol. 54, no. 6, 2017, pp. 786-801.

[12] W. Wang, K. Zhu, H. Wang, & Y.-C. J. Wu, “The impact of sentiment orientations on successful crowdfunding campaigns through text analytics,” vol. 11, no. 5, 2017, pp. 229-238.

[13] Q. Du, Z. Qiao, W. Fan, M. Zhou, X. Zhang, & G. Wang, “Money Talks :A Predictive Model on Crowdfunding Success Using Project Description,” Twenty-first Americas Conference on Information Systems, AMCIS, 2015, pp. 1–8.

[14] M. Schuckert, X. Liu, & R. Law, “A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently,” International Journal of Hospitality Management, vol. 48, 2015, pp. 143-149.

[15] M. Dash & H. Liu, “Feature selection for classification,” Intelligent data analysis, vol. 1, no. 3, 1997, pp. 131-156.

[16] O. Frank, B. Brors, A. Fabarius, L. Li, M. Haak, S. Merk, U. Schwindel, C. Zheng, M. C. Müller, N. Gretz, R. Hehlmann, A. Hochhaus & W. Seifarth, “Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients,” Leukemia, vol. 20, no. 8, 2006, pp. 1400-1407.

[17] R. Stoean & C. Stoean, “Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection,” Expert Systems with Applications, vol. 40, no. 7, 2013, pp. 2677-2686.

[18] W. Liu, J. Zhai, H. Ding, & X. He, “The research of algorithm for protein subcellular localization prediction based on SVM-RFE,” 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI), IEEE, 2017, pp. 1-6.

[19] Z. Shao, S. Yang, F. Gao, K. Zhou & P. Lin, “A new electricity price prediction strategy using mutual information-based SVM-RFE classification,” Renewable and Sustainable Energy Reviews, vol. 70, 2017, pp. 330-341.

[20] C. Cortes & V. Vapnik, “Support-Vector Networks,” Machine Leaning, vol. 20, no. 3, Sept. 1995, pp. 273-297.

[21] A. Sun, E.-P. Lim, & Y. Liu, “On strategies for imbalanced text classification using SVM: A comparative study,” Dicision Support Systems, vol. 48, no. 1, 2009, pp. 191-201.

[22] Y. Xi, “Chinese review spam classification using machine learning method,” 2012 International Conference on Control Engineering and Communication Technology, ICCECT, 2012, pp. 669-672.

[23] M. Ott, Y. Choi, C. Cardie, & J.T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT, 2011, pp. 309-319.

[24] J. Meng, H. Lin, & Y. Yu, “A two-stage feature selection method for text categorization,” Computers and Mathematics with Applications, vol. 62, no. 7, 2011, pp. 2793-2800.