The key successful factors of video and mobile game crowdfunding projects using a lexicon-based feature selection approach

Mu-Yen Chen[1] · Jing-Rong Chang[2] · Long-Sheng Chen[2] · En-Li Shen[2]

Abstract
The emergence of crowdfunding has given many capital demanders a new fund-raising channel, but the overall project success rate is very low. Many scholars have begun to discover key successful factors of crowdfunding projects. Previous studies have used questionnaires survey to identify important project features. In addition to requiring a lot of manpower and time, there may also be sampling bias. Moreover, related studies also reported that the project description will affect the success of the crowdfunding project, but there is no research to tell fundraisers which success factors should be included in the content of the project description. Besides, in recent years, game crowdfunding projects have been attracted lots of attention in terms of total fundraising amount and number of projects. Moreover, in traditional feature selection and text mining approaches, the selected terms are un-organized and hard to be explained. Therefore, this study will focus on real video and mobile game project descriptions to replace conventional questionnaires. To solve these issues, we present a lexicon-based feature selection method. We attempt to define "content features" and build lexicons to determine the attributes' values. Three feature selection methods including decision tree (DT), Least Absolute Shrinkage and Selection Operator (LASSO), and support vector machine–recursive feature elimination (SVM–RFE) will be employed to find organized candidate key successful factors. Then, support vector machines (SVM) will be used to evaluate the performances of candidate feature subsets. Finally, this study has identified the key successful factors for video and mobile games, respectively. Based on the experimental results, we can give fundraisers some useful suggestions to improve the success rate of crowdfunding projects.

Keywords Crowdfunding · Text mining · Feature selection · Key succesful factors · Games

1 Introduction

With the rapid development of the Internet, crowdfunding has emerged accordingly (Mollick 2014). Zhang and Chen (2019) defined crowdfunding as a fundraising method in which individuals, groups or companies raise funds from the public through the Internet (Cox et al. 2018; Kim et al. 2017). It's reported that the total amount of funds raised has exceeded US\$5.2 billion, and the total number of participants has exceeded 18.47 million in Kickstarter (Kickstarter. com August 2020). Crowdfunding brings a totally new channel of fundraising for straterups which often encounter insufficient resources in the early stage (Cosh et al. 2009; Kim et al. 2017). Different from traditional fundraising channels, fundraisers even could build relationships with investors on the platform, and eventually become partners or customers. Moreover, entrepreneurs can also use crowdfunding to attract customers to participate in the development of new products to improve products and better meet consumer needs (Petruzzelli et al. 2019).

But it's reported that the success rate of crowdfunding is very low. According to statistics of Kickstarter on August 2020, the overall success rate is only about 37.97%. It can be seen that initiating a crowdfunding project is not easy

✉ Long-Sheng Chen
lschen@cyut.edu.tw

Mu-Yen Chen
mychen119@gs.ncku.edu.tw

Jing-Rong Chang
chrischang@cyut.edu.tw

En-Li Shen
s10714616@gm.cyut.edu.tw

[1] Department of Engineering Science, National Cheng Kung University, Tainan City 701401, Taiwan

[2] Department of Information Management, Chaoyang University of Technology, 168, Jifeng E. Rd., Wufeng District, Taichung 413310, Taiwan

to succeed. Consqently, how to improve the success rate of crowdfunding projects is one of major concerns of all fundraisers. To achieve this goal, lots of researchers have paid attention on this issue. For instances, Liang et al. (2020) found that the number of pictures, the number of videos, the number of comments, and the number of updates have a positive impact on the success of crowdfunding, but the readability is negative. Fernandez-Blanco et al. (2020) found that comments and updates are helpful to the development of crowdfunding projects. Wang et al. (2017) believe that the text content of the crowdfunding project description will affect the decision-making of supporters. Lagazio and Querci (2018) detailed text content is more effective than introduction videos. Zhou et al. (2018) found that the length, readability, semantics of the project description and the fundraiser's past fundraising experience will affect the success of fundraising. Kim et al. (2017) stated that detailed project descriptions have a positive impact on the success of fundraising projects. These studies reported that the project description will affect the success of the crowdfunding project, but there is no research to tell fundraisers which success factors should be included in the content of the project description.

Moreover, related works usually use qualitative research methods with data surveyed by questionnaires. In addition to requiring a lot of manpower and time, there may also be sampling bias. In recent years, text minig and feature selection have been succesfully applied to deal with the huge text reviews in social media (Wang et al. 2017). Feature selection mainly processes numerical data, although the natural language processing technology (NLP) in text mining technology can process text data. But, if we only use both feature selection and text mining approaches, we cannot obtain organized factors. The selected terms will be quite difficult to be explained, and the number of extracted features will be very huge (Chen et al. 2015).

Besides, in terms of total fundraising amount and number of projects, game projects have been ranked as top one in crowdfunding platdforms. Unfortunately, only few existing literatures focus on the research of game crowdfunding. To solve this problem, we present a lexicon-based feature selection method which aims to discover the crucial content features of project descriptions toward "video game" and "mobile game".

To sum up, this study will collect real video games and mobile games projects as our research data. In our proposed method, we attempt to define "content features" that may affect the success of game crowfunding projects based on literature review, and establishing lexicons to determine the value of features. The natural language processing (NLP) technology in text mining can be used to process collect text data. Then, three feature selection methods including decision trees (DT), Least Absolute Shrinkage and Selection Operator (LASSO), support vector machine–recursive feature elimination (SVM–RFE) will be employed to select candidate key successful factors, and finally support vector machines (SVM) will be performed to evaluate the performance of candidate feature subsets. The discovered key success factors can provide fundraisers with a basis when establishing crowdfunding projects to help them increase the success rate of video game and mobile game projects.

2 Literature review

2.1 Crowdfunding

Crowdfunding is an innovative fundraising method for many entrepreneurs (Yang et al. 2020; Borrero-Dominguez et al. 2020). Traditional fundraising methods are very difficult for most of startups because they cannot propose equivalent values (Ramadani 2009). So, investors cannot assess risks (Schwienbacher and Larralde 2010). Unlike traditional fundraising channels, crowdfunding has lower costs and is very suitable for startups, non-profit organizations and individuals (Vanacker and Manigart 2010; Liang et al. 2020). Now, crowdfunding has become one of the most important fundraising methods (Cumming et al. 2020).

In crowdfunding, fundraisers initiate projects on platforms, express their products, services or ideas through projects, and use the power of the Internet to obtain funds (Ziegler et al. 2018; Hollas 2013; Colgren 2014; Mollick 2014; Mollick and Robb 2016). Gerber et al. (2012) believe that the use of crowdfunding can also build relationships and form a social network. While sponsors invest to obtain income or return, some sponsors wanted to gain recognition from others (Bretschneider et al. 2014). Results of crowdfunding could be considered as an indicator of a new product or service before entering the market (Mollick 2014; Meyskens and Bird 2015).

Most of researches aim to discover the success or failure factors of crowdfunding projects (Short et al. 2017), but there are also some studies that explore the contribution behavior and motivation of supporters (Xu et al. 2016; Cox et al. 2018; Shneor and Munim 2019). Other scholars have studied game crowdfunding projects. For examples, Lax (2017) interviewed crowdfunding creators in three different game industries. He found six success factors, including project rewards, project goals, the quality of products and projects, project team, fundraiser information and project preparation. Smith (2015) discussed the fundraising process of game fundraising projects and observed common occurrences between developers and players to study how interaction affects product development and production. Some published works (Wang et al. 2017; Kim et al. 2017; Lagazio and Querci 2018; Zhou et al. 2018) have reported

that the project description is important for the success of the crowdfunding project, but no available studies can tell fundraisers which success factors should be included in the content of the project description. Therefore, this study will use the real game crowdfunding project for experiments and analysis, and try to discover key content features of game crowdfunding projects.

2.2 Factors of games

From the available literature, relatively few studies focus on game crowdfunding. In order to find the key successful factors of game crowdfunding projects, this study attempts to find the potential factors related to games from literature. For examples, Caci et al. (2018) analyzed players who play "Pokémon GO", discussed the interaction between players' play motivation, player's personality, and play habits, and conducted research on individual differences in play motivation and personality characteristics. The work of Chan et al. (2017) shows the information, interactivity, and entertainments of mobile games have a positive impact on the trust and loyalty of players. Hsiao and Chen (2016) identified the factors that affect the perceived value and loyalty of games, including fun, flexibility of play time, interactivity, and rich rewards. Chen et al. (2010) defined 19 online game quality elements for massively multiplayer online role-playing games. Zhao and Fang (2009) showed that online game technical factors (game stories, game images, game duration, game control, and the quality of game service) has a significant impact on the game fun of players; game fun and social norms have a positive impact on game intentions; social norms, the quality of online game communities and game intentions are important predictive indicators for online game loyalty. Thereofre, we used results of literature review to build content features.

2.3 Text mining

Text mining can be regarded as a process of editing, organizing, and analyzing a large number of documents. Its main task is to convert text into data for analysis through language analysis and natural language processing (Dreisbach et al. 2019). In this era of information explosion, these large amounts of unstructured or semi-structured text data need to be processed with text mining technology to discover hidden knowledge (Thomaz et al. 2017). Miner et al. (2012) defined the main procedures of text mining as data retrieval and processing, word segmentation, feature selection, classification and clustering, text representation and interpretation.

Nowadays, lots of studies use text mining to find the potential characteristics in the introduction of crowdfunding projects. For examples, Wang et al. (2017) analyze the text content of crowdfunding project descriptions and the emotions of the project founders when making the project descriptions. Du et al. (2015) studied the quality and source credibility of the description of crowdfunding projects, and analyzed the impact on the success of crowdfunding projects. Text mining is also used in various areas. Loureiro et al. (2020) used text mining technology to analyze the full text of VR and AR-related journals and conference papers. Zhong et al. (2020) combined deep learning and text mining to automatically analyze the hazard records of building construction.

Moreover, because traditional questionnaire surveys are prone to experimental effects and the information brought by online text content will be more objective, massive, and less sample biased than using questionnaires (Schuckert et al. 2015). As a result, this study will use text mining methods to process the unstructured text content of crowdfunding projects. But, if we only use text mining approaches which acquire knowledge form term-document matrix (TDM), we cannot obtain organized factors, and the number of extracted features will be very huge (Chen et al. 2015). Therefore, this study proposed a lexicon-based feature selection method which uses NLP and lexicons to construct content features from project descriptions.

2.4 Feature selection

Feature selection aims to find important features from original feature set (Zhao et al. 2020). By using selected crucial features, it can improve the prediction accuracy of the classifier and reduce the training time (Devi and Sabrigiriraj 2018; Dash and Liu 1997). Therefore, this study will use three feature selection methods to find the important factors that will affect the success of crowdfunding from the project features and content features of crowdfunding.

2.4.1 Decision trees

DT is an easy-to-understand classification method. It is not only easy to use but also can quickly find rule conditions for high-dimensional data (Yang 2019; Ma et al. 2016; Sok et al. 2016). In this study, the leaf nodes of rule tree will be considered as important factors. Being one of famous feature selection methods, DT has been successfully applied in various fields. For examples, Kwon et al. (2020) used decision trees to determine important factors which would affect the shopping behavior of second-hand stores. Namazkhan et al. (2020) used decision trees to analyze key factors affecting household natural gas consumption. Chen et al. (2021) employed decision trees to select the crucial factors of increasing customers' satisfaction from online customer reviews. Consequently, this study will use DT as one of the feature selection methods for crowdfunding projects.

2.4.2 LASSO

LASSO is a regression analysis method that can perform feature selection and regularization at the same time. The ultimate goal of reducing the variables to zero is to obtain the feature subset that minimizes the prediction error of the variable (Tibshirani 1996). LASSO obtains a more refined model by establishing a valve function, and reduces the sum of coefficients by the square of the least square method, and compresses the sum of absolute values of coefficients to less than the constant 1 (Wang et al. 2018).

LASSO has been suceesfully applied in various fields. For examples, McEligot et al. (2020) used LASSO to determine the most relevant variables that cause breast cancer. Schmidt et al. (2020) utilized LASSO to detect and identify global position system (GPS) fraud. Guenther and Sawodny (2019) employed LASSO to determine the important factors affecting the temperature comfort of open offices in Singapore. Therefore, this study will use SVM–RFE as one of the feature selection methods.

2.4.3 SVM–RFE

SVM–RFE is another famous feature selection technique. It can find relatively important features from a large number of features. Through the weight vector obtained during training, all the features are arranged in descending order, and the feature with the smallest coefficient is deleted in each down generation, and then retrained and sorted. Repeat the above steps to the end. We can get all the features in descending order (Frank et al. 2006; Witten et al. 2011). SVM–RFE has been successfully applied in many areas. For examples, Wang et al. (2019) used SVM–RFE to identify autism. Shao et al. (2017) applied SVM–RFE to predict the price of electricity in electricity market analysis. In the work of Liu et al. (2017), they used SVM–RFE to locate protein sub-cells in biomedicine. Chang et al. (2020) applied SVM–RFE as one of the feature selection methods to determine the important factors of influencing O2O trust (Chang et al. 2020). As a result, this study will use SVM–RFE as one of the feature selection methods.

2.5 SVM

SVM (Cortes and Vapnik 1995) is a supervised machine learning model used for classification and regression analysis. Because SVM can be applied to various linear and non-linear classification problems, and has a complete theoretical framework and tools, it has been widely used (Paul et al. 2016). For examples, Gamal et al. (2019) utilized SVM with other a variety of classifiers to perform sentiment analysis and classification of Arab tweets and comments on twitter.

Vijayakumar and Muhammad (2019) employed SVM, NB and maximum entropy with natural language processing methods to indentify sapm comments on the online forum. Krammer et al. (2019) utilized SVM, RBF, and MLP to analyze online comments for abnormal behavior. To sum up, SVM has an excellent performance on text classification, so this study will use SVM to evaluate the performance of selected candidate feature subset.

3 Proposed methodology

This section will introduce the implementation procedure of the proposed lexicon-based feature selection method. The procedure shown in Fig. 1 can be divided into nine steps. First, we collect data from a real crowdfunding website. Next, the "project features" and "content features" have been induced from the relevant literature. In the content features, we focus on the aspect of the game's success. Then, we set up the lexicons for content features, and compute attributes' value based on built lexicons. Later, feature selection methods will be utilized to find the candidate feature subsets, and finally the SVM classifier has been performed to evaluate the performance of the selected feature subsets. After that, we
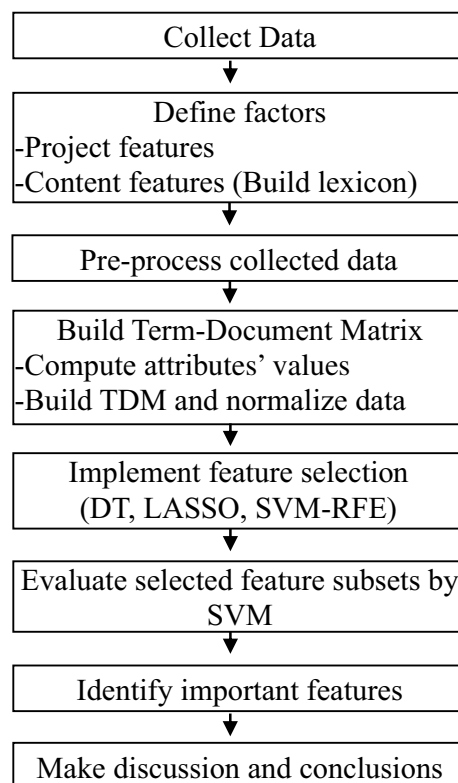

Fig. 1 The implemental procedure of proposed method

can identify important features, and make specific suggestions accordingly. The following describes the steps of the method used in this study.

Step 1: Collect data

We use Kickstarter (https://www.kickstarter.com), one of the most popular crowdfunding platforms in the world, as the data sources. The scope of the collection is the "video game" project and "mobile game" under the "game" category in the platform. Therefore, this research will be divided into experiment #1 (video game projects) and experiment #2 (mobile game projects).

Python language will be employed to write a web crawler tool. The crawled data is divided into two parts. The first part includes content features which will capture the content of project descriptions. The second part involves project features, including the target amount, video clip, number of updates, number of comments, project duration, fundraising experience, number of pictures, description length, and number of investors.

Step 2: Define factors

This study divides the features into two parts, project features and content features. The project features will be induced from related literature of crowdfunding projects. The content features are from the literature related to the game theme to find out the aspects and words related to the success of the game. For content features, we will gather feature related keywords to build lexicons. In one lexicon, the keywords' similar words, synonyms, and antonyms also will be included in the lexicon. Based on constructed lexicons,

we can calculate the values of features. We'll clarify how to define project fetaures and content feature, respectively, as follows.

Step 2.1 Define project features

According to relevant literature on crowdfunding, we find out project features that may affect the success of crowdfunding projects, and give clear definitions, since these features have not been used for game crowdfunding projects. The definitions of project features have been shown in Table 1.

Step 2.2 Define content features

The lexicons of content features will be constructed from game-related literature. From these published works, we collect rellant keywords, factors, and dimensions, then organize them into content features and representative words. Next, we use the website (https://www.thesaurus.com/) to find similar words, synonyms, antonyms and related vocabularies of these representative words to build a lexicon. The defined content features and their lexicons have been shown in Table 2. After these lexicons are established, it will be compared with the text of the project description, and the frequency of the words in the lexicon will be calculated, and the word frequency will be defined as the values of content features.

Step 3: Pre-process collected data

The captured project descriptions belong to unstructured text data, so they need to be pre-processed before it can be compared with the constructed lexicons. The pre-processing steps of text data are as follows:

Table 1 Project features

Notation	Project features	Definition	Attribute value	Supports
X1	Target amount	The fundraising goal set when the project is established	Continous (USD)	Cholakova and Clarysse (2015) and Li et al. (2018)
X2	Vedio clip	The project introduction contains a video or not	Binary (1: YES/0: No)	Lagazio and Querci (2018), Li et al. (2018) and Liang et al. (2020)
X3	Updates	The number of all updates at the end of the project	Continous	Lagazio and Querci (2018), Kim et al. (2017) and Fernandez-Blanco et al. (2020)
X4	Number of comments	The number of comments includes questions from investors and answers from fundraisers	Continous	
X5	Duration	The number of days from project starts to end	Continous (days)	
X6	Fundraising experiences	The number of fundraising projects previously owned by the fundraiser	Continous	Zhou et al. (2018) and Kim et al. (2017)
X7	Number of pictures	The number of pictures used in the project description	Continous	Li et al. (2018) and Liang et al. (2020)
X8	Length of description	The number of words/characters in the description of the project	Continous	Zhou et al. (2018)
X9	Number of investors	The total number of investors	Continous	Li et al. (2018)

Table 2 Content features and their lexicons

Notation	Content features	Examples of lexicons	Supports
X10	Contents	Background, story, task, backdrop, culture, education, environment, history…	Chen et al. (2010) and Zhao and Fang (2009)
X11	Interaction	Guild, chat, interactive, communication, conflict, federation, club, company, corporation, group…	Chen et al. (2010), Hsiao and Chen (2016) and Chan et al. (2017)
X12	Role	Role, characteristic, difference, ability, act, appearance, aspect, character, part, performance…	Chen et al. (2010)
X13	System	Error, update, aegis, repair, server, hardware, equipment, GM, recourse, OS, adjustment, improvement…	Chen et al. (2010) and Zhao and Fang (2009)
X14	Motivation	Evaluation, reputation, pleasure, win, achievement, peer, character, fame, honor, influence…	Caci et al. (2018)
X15	Promotion	Promotion, discount, merch, alliance, television, network, magazine, advertisement, trailer, advancement, Facebook, Youtube…	Chen et al. (2010)
X16	Screen	Screen, scenes, map, background, character, pattern, animation, cover, curtain, 2D, 2.5D, 3D…	Chen et al. (2010) and Zhao and Fang (2009)
X17	Sound effects	Music, sound, melody, opera, piece, rap, rock, singing, soul, tune, acoustic…	Chen et al. (2010)
X18	Operation	Control, safety, ergonomics, authority, curb, discipline, domination, force, government, jurisdiction, management…	
X19	Price	Price, cheap, expensive, amount, bill, cost, demand, discount, estimate, expenditure, expense…	
X20	Appearance	Exterior, pack, coating, facade, skin, surface, appearance, aspect, cover, covering…	

Step 3.1: Delete meaningless words, characters, or icons

We delete non-English text, special characters and emoticons.

Step 3.2: Segment sentences

We use uni-gram for word segmentation. The uni-gram method uses a word as a unit to segment the sentence. Take "Nice to meet you" for example. After uni-gram segmentation, we will get "Nice", "to", "meet" and "you".

Step 3.3: Remove stop words

This study uses the stop word list from the Snowball website (http://snowball.tartarus.org/algorithms/english/stop.txt) to delete stop words, such as and, or, with.

Step 4: Build term-document matrix

This step is divided into two parts. The first part is to calculate the word frequency of the built corpus to be the values of content features. The second part is to establish the Term-Document Matrix (TDM). In this study, we use content features, project features, and all features to build TDM, respectively.

Step 4.1: Compute attributes' values of content features

This study uses QDA miner to deal with textual data. We compare the constructed lexicons with the collected data, calculate occurrence frequency of words in lexicons, and use the frequency as the values of a content feature. After calculating the word frequency of all content features, we discretize all content features, and define the value above the average value as 1. Otherwise, the value below the average value will be defined as 0.

Step 4.2: Build TDM and normalize data

We use project features and content features to build TDM. The output value is Y. If Y equals to 1. It means the project is successful (the raised funds exceed the target amount). If Y equals to − 1. It means the project is failed. Then, all colleted data will be normalized into the interval [0, 1] according to Eq. (1). In this equation, $X'_{n,m}$ is the value to be normalized for the document d_m in the feature X_n. $X'_{n,min}$ is the minimum value of all documents under the feature X_n. $X'_{n,max}$ is the maximum value and $X_{n,m}$ is the normalized value of the data.

$$X_{n,m} = \frac{X'_{n,m} - X'_{n,min}}{X'_{n,max} - X'_{n,min}}. \tag{1}$$

After normalization, we sort all the data randomly, and then we implement fivefold cross-validation (CV) experiement which divides all data into 5 equal parts. 4 of them are used as training set, and the other part is as a test set in turn.

Step 5: Implement feature selection

We use DT, LASSO, SVM–RFE to select key features. Since we implement fivefold CV experiment, we'll have five

results of feature selection for each method. And we will pick the features whose occurrence frequency is larger than 3 to be our candidate feature subsets.

Step 5.1: Decision tree

This study will use the C5.0 algorithm to build a decision tree. This method is a kind of intuitive and highly efficient supervised machine learning model which is suitable for classification and regression. It takes up less system resources and memory. Compared with other machine learning methods, execution speed is its major advantage. When used in feature selection, we treat the leaf nodes of the rule tree as an important factor. The steps to build a decision tree are as follows.

(1) Define the input (X) and output (Y) variables of the decision tree.

(2) Establish a decision tree using C5.0 algorithm.

(2.1) Establish an initial rule tree.

(2.2) Prune this rule tree.

(2.3) Select the best-performing rule tree and leaf nodes.

Step 5.2: LASSO

LASSO can perform feature selection and normalization at the same time. The main reference formula is shown in Eq. (2), where λ is the adjustment coefficient, K is the number of data, and x is the explanatory variable.

$$min \sum_{k=1}^{K} (y_k - \beta_0 - x_{1,k}\beta_1 - \cdots - x_{g,k}\beta_g)^2, \quad s.t. \sum_{i=1}^{g} |\beta_i| \leq \lambda. \tag{2}$$

According to Eq. (2), it is found that the regression parameter value βj will be restricted by a specific penalty selection criterion, and then an appropriate variable can be selected. Given g explanatory variables, the value of λ will affect the parameter $\hat{\beta}$ estimation. When the λ value approaches infinity, the parameter $\hat{\beta}$ estimation value will not be restricted by Eq. (2). On the contrary, when the λ value is adjusted to 0, all parameter estimation values will be zero.

When the value of λ is gradually enlarged from 0, the explanatory variable x that is strongly correlated with y_k will change and differ from 0, which means that the explanatory variable is important. Therefore, in this experiment, the basis for identifying important features is to check whether the coefficient is 0, and the coefficient is not 0 to represent an important factor. Perform LASSO on all fivefolds, then extract features that are not zero from the results as important factors, and calculate features that occur more than or equal to 3 times to build a feature subset.

Step 5.3 SVM–RFE

The detailed SVM–RFE (Chang et al. [5]) algorithm has been shown as bellow.

Inputs:

Training examples

$X_0 = [x_1, x_2, ...x_k, ... x_l]^T$

Class labels

$y = [y_1, y_2, ...y_k, ... y_l]^T$

Initialize:

Subset of surviving features

$s = [1, 2, ... n]$

Feature rank list

$r = [\ ]$

Repeat until $s = [\ ]$

Restrict training examples to good feature indices

$X = X_0(:, s)$

Train the classifier

$\alpha = S V M - train\ (X, y)$

Compute the weight vector of dimension length(s)

$w = \sum_k \alpha_k y_k x_k$

Compute the ranking criteria

$c_i = (w_i)^2, for\ all\ i$

Find the feature with smallest ranking criterion

$f = arg\ min\ (c)$

Update feature rank list

$r = [s(f), r]$

Eliminate the feature with smallest ranking criterion

$s = s(1 : f - 1, f + 1 : length(s))$

Output:

Feature ranked list r.

This research uses fivefold CV experiment. In each fold, we execute SVM–RFE once. The sorted feature set of fivefold will be obtained. We extract features whose weights are larger than the median. Then, we will calculate the features with a frequency greater than or equal to 3 times to build a feature subset.

Step 6: Evaluate the performance of selected feature subsets by SVM

This study will use the SVM classifier to evaluate the original feature set and the candidate feature subset selected in step 5. In this study, we use the radial basis function (RBF) kernel for building the SVM classifier. The measurement indicators include positive accuracy (PA), negative accuracy (NA), geometric mean (G-Mean, GM) of PA and NA, overall accuracy (OA), F1-measure (F1), training time of building classifiers. If the feature candidate subset is built with a smaller number of features, the classifier can have almost or even better classification

performance, which means that the extracted feature subset contains a higher amount of information, so it is regarded as a few important features.

Step 7: Identify important features

In this step, the important features will be identified based on the evaluation results of step 6. In addition to confirming important content features, we will also confirm the impact on content features after adding project features. In other words, we will confirm the performance of feature selection based on content features, project features, and all features, respectively. Then, we can identify the crucial content features and project features that affect the success rate of video game and mobile game projects, individually.

Step 8: Make discussion and conclusions

For the determined important features, we will carry out statistical analysis on every single one feature, and discover the relationship between these selected important features and fundraising success. Finally, based on the experimental results, this research will give fundraisers suggestions that will help improve the success rate of video games and mobile games crowdfunding projects.

4 Results

4.1 Data collection and prepreparation

This study uses the "Python" programming language to write a web crawler program, and extract project introduction text and project features. The data source comes from projects of "Video Game" and "Mobile Game" under the game category on the Kickstarter platform. The crawled data is the text content of the project description for 11 content features. The project content description example is shown in Fig. 2. In addition, 9 project features are also collected.

In video game projects, the project achievement rate is greater than or equal to 100% is labeled as a successful project, and the project achievement rate is less than 100% is marked as a failed project. The total number of video game projects is 14,495, because the total number of projects exceeds 10,000, and due to restrictions on the Kickstarter platform, many projects are no longer crawled. So, this study collected one-tenth of the total number of video game projects as our experimental data. From 2018 to 2020, a total of 1400 projects were collected, 700 successful projects, and 700 failed projects.

In mobile game projects, the project achievement rate is greater than or equal to 100% is marked as a successful project, and the project achievement rate is less than 100% is labeled as a failed project. The total number of mobile game projects is 3642. This study has collected all mobile game projects, with 336 successful projects and 3306 failed projects, from 2013 to 2020. The descriptive statistics of the collected data are shown in Tables 3 and 4.

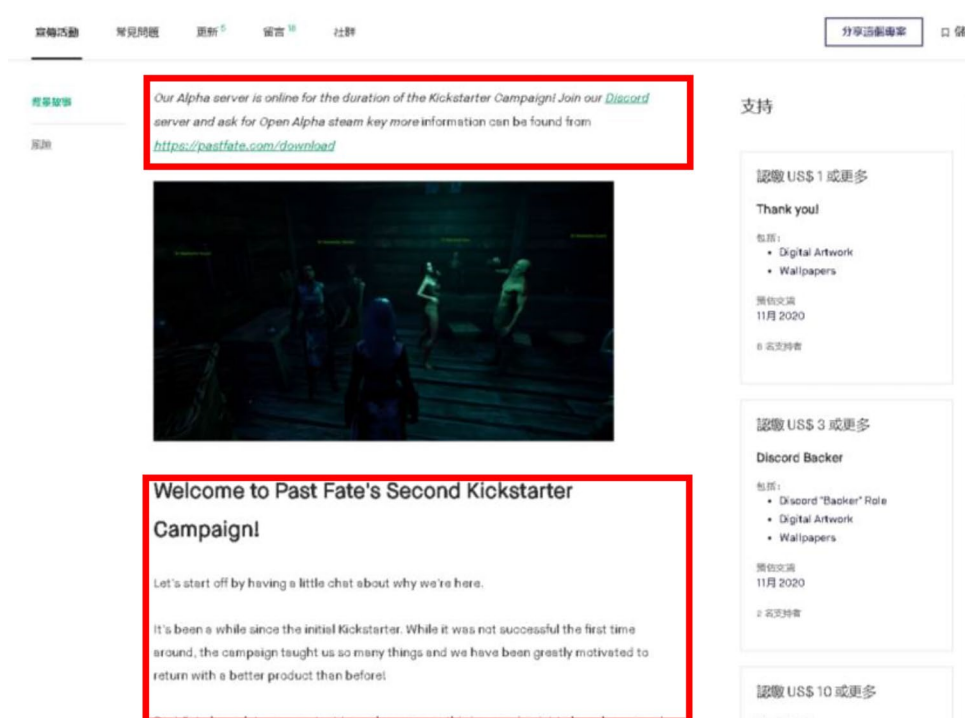**Fig. 2** An example of project content description in Kickstarter.com

Table 3 Descriptive statistics of video game projects

Features		Successful (700 projects)			Failed (700 projects)		
		Max	Min	Mean	Max	Min	Mean
Project features	Target amount (USD)	1,285,000	1	17,891.76	1,230,399	10	31,463.09
	Vedio clip	1	0	–	1	0	–
	Updates	125	0	17.55	55	0	2.53
	Number of comments	19,328	0	179.62	93	0	3.17
	Duration	60	6	32.20	229	2	38.08
	Fundraising experiences	34	0	1.32	17	0	0.32
	Number of pictures	116	0	27.68	101	0	14.05
	Length of description	4251	0	1191.62	5223	0	790.64
	Number of investors	58,730	2	943.92	833	1	30.72
Content features	Contents	61	0	15.74	42	0	11.15
	Interaction	21	0	5.65	19	0	3.74
	Role	34	0	11.67	39	0	7.97
	System	21	0	6.43	21	0	4.28
	Motivation	26	0	7.58	23	0	5.72
	Promotion	26	0	7.08	24	0	4.57
	Screen	39	0	10.17	32	0	6.61
	Sound effects	9	0	1.93	8	0	1.16
	Operation	21	0	2.50	9	0	1.31
	Price	14	0	2.96	12	0	1.97
	Appearance	45	0	5.07	12	0	2.44

Table 4 Descriptive statistics of mobile game projects

Features		Successful (336 projects)			Failed (3306 projects)		
		Max	Min	Mean	Max	Min	Mean
Project features	Target amount (USD)	500,000	1	11,110.75	1,000,000	1	29,707.69
	Vedio clip	1	0	–	1	0	–
	Updates	100	0	13.98	43	0	1.32
	Number of comments	5765	0	56.43	403	0	1.56
	Duration	70	7	32.56	89	1	33.89
	Fundraising experiences	8	0	0.26	12	0	0.16
	Number of pictures	98	0	8.58	92	0	6.74
	Length of description	4627	0	896.63	5275	0	601.45
	Number of investors	14,081	1	254.54	1023	0	16.49
Content features	Contents	36	0	12.09	42	0	8.74
	Interaction	19	0	4.39	22	0	3.02
	Role	33	0	9.09	35	0	6.14
	System	18	0	5.41	19	0	3.81
	Motivation	24	0	6.44	22	0	4.97
	Promotion	24	0	5.61	29	0	3.82
	Screen	29	0	7.51	32	0	5.22
	Sound effects	7	0	1.50	9	0	0.87
	Operation	20	0	1.86	21	0	1.17
	Price	11	0	2.35	12	0	1.65
	Appearance	38	0	4.11	46	0	2.7

Table 5 Examples of pre-process collected data

Before	After
Delete meaningless characters or icons	
Hello world! I'm so excited that you're here!!!	Hello world I'm so excited that you're here
Remove stop words	
Thank you for joining us at our Developer Commentary Livestream	Thank joining Developer Commentary livestream

4.2 Data preprocessing

For raw data, we firstly delete special characters and non-English words in the text to avoid errors in subsequent experiments. The package software "QDA Miner" has been employed to segment sentences (unigram) and then remove stop words. Examples of non-English word deletion and removal stop words are shown in Table 5.

Next, we use the built lexicons to calculate the word frequency as values of content features and then discretize the content features into $+1$ (larger or equals to mean value) and -1 (below mean value). Then, we normalize data into [0, 1] and build TDMs of content features, project features, and all features, respectively.

4.3 Feature selection

This study uses three feature selection methods to select features for "all features", "project features" and "content features", respectively. DT uses the C 5.0 commercial version "See 5", LASSO uses the "Matlab R2013b" to program, and SVM–RFE uses "Weka 3.8" to find important features and build candidate feature subsets.

See 5 uses the default "global pruning" and "pruning CF $= 25\%$". The parameter alpha value of LASSO is set to 1, and the λ value is preset to calculate the maximum λ value built in Matlab, and the ratio of the maximum value to the minimum value in the sequence is found with the Lambda Ratio function to calculate the value ratio of λ using CV $= 5$ (fivefold cross-validation method) to find the optimal λ value. In SVM–RFE, the attribute evaluator parameter selects "ReliefFattributeEval", and the other parameters are default values. The results of experiment

Table 6 Summary of feature selection results

Experiment	Method	Features	Candidate feature subsets	Extracted features
#1 (Video game)	DT	All	DT1	X1, X3, X9
		Project	DT2	X1, X3, X4, X9
			DT3	X1, X3
		Content	DT4	X11, X12, X13, X15, X16, X17, X18, X19, X20
			DT5	X11, X17, X18, X19, X20
	LASSO	All	LASSO1	X1, X3, X5, X6, X7, X20
		Project	LASSO2	X1, X3, X5, X6, X7
		Content	LASSO3	X13, X15, X16, X17, X20
			LASSO4	X15, X17, X20
	SVM-RFE	All	SVM-RFE1	X3, X7, X10, X11, X12, X13, X15, X16, X19, X20
		Project	SVM-RFE2	X2, X3, X5, X6, X9
		Content	SVM-RFE3	X10, X11, X15, X17, X20
			SVM-RFE4	X11
#2 (Mobile game)	DT	All	DT#1	X1, X3, X9
			DT#2	X1, X3
		Project	DT#3	X1, X3
			DT#4	X1
		Content	–	–
	LASSO	All	LASSO#1	X1, X3
		Project	LASSO#2	X3
		Content	LASSO#3	X12, X13, X15
	SVM-RFE	All	SVM-RFE#1	X3, X5, X7, X9, X10, X11, X14, X15, X16, X19
		Project	SVM-RFE#2	X3, X5, X6, X7, X8
		Content	SVM-RFE#3	X10, X13, X15, X17, X18, X19
			SVM-RFE#4	X15, X17, X19

#1 (video game project) and #2 (mobile game project) are shown in Table 6.

4.4 Evaluated by SVM

This study uses the SVM algorithm as the classifier, using the RBF kernel function and C-SVM classification in "LIB-SVM". The parameter settings use the grid.py tool which is built on LIBSVM to find the optimal parameter settings. These optimal parameters will be employed to buid SVM classifiers to perform performance evaluation, and find the best feature subset.

4.4.1 Results of experiment #1 (video game)

Table 7 summarizes the evaluation results of "all features" of video game projects. According to Table 7, it can be seen that the DT1 feature subset is superior to the original features and other feature subsets on all metrics, and the overall accuracy (OA) reaches 97.64%. Therefore, the DT1 feature set is selected as the best feature subset for "all features".

The performance evaluation results of "project features" have been shown in Table 8. From this table, it can be seen that all the metrics of the DT2 feature subset and the DT3

feature subset are very similar. Both DT2 and DT3 are better than the original and other feature subsets. Their overall accuracies reach 97.49% and 97.79%, while the DT3 feature subset contains only 2 features less than 4 features in the DT2 feature set. Fewer features can show the same performance. Therefore, the DT3 feature subset is selected as the best feature subset for "project features".

The performance evaluation results of "content features" are shown in Table 9. From this table, we can see that the overall accuracy (OA) of the DT5 feature set and the LASSO4 feature set reached 64.21% and 64.07%, respectively. The F1 index reached 63.13% and 63.05%, respectively. And the training time was 0.19 s and 0.05 s. But LASSO4 only has three features. Relatively speaking, a small number of features can show the same performance. This means that these three features are more important. Therefore, the LASSO4 feature set is selected as the best feature set for "content feature".

4.4.2 Results of experiment #2 (mobile game)

Table 10 shows the performance evaluation results of "all features". From this table, it can be seen that all the indicators of the DT#1 feature set and the DT#2 feature set are

Table 7 Evaluation results of all features in video game projects

Feature sets Index	Original (20 features) Mean (SD)	DT1 (3 features) Mean (SD)	LASSO1 (6 features) Mean (SD)	SVM-RFE1 (10 features) Mean (SD)
PA (%)	89.82 (2.50)	98.02 (1.11)	92.51 (2.95)	89.18 (2.36)
NA (%)	84.02 (2.35)	97.37 (2.41)	97.22 (2.06)	80.20 (2.71)
GM (%)	86.86 (1.91)	97.68 (0.75)	94.81 (1.11)	84.56 (2.24)
OA (%)	86.64 (1.95)	97.64 (0.78)	94.64 (1.31)	84.07 (2.32)
F1 (%)	86.08 (2.12)	97.62 (0.83)	94.79 (1.18)	82.94 (2.64)
Time (s)	0.19 (0.29)	0.06 (0.02)	0.35 (0.13)	0.08 (0.03)
Extracted features	X1–X20	X1, X3, X9	X1, X3, X5, X6, X7, X20	X3, X7, X10, X11, X12, X13, X15, X16, X19, X20

Table 8 Evaluation results of project features in video game projects

Feature sets Index	Original (9 features) Mean (SD)	DT2 (4 features) Mean (SD)	DT3 (2 features) Mean (SD)	LASSO 2 (5 features) Mean (SD)	SVM-RFE2 (5 features) Mean (SD)
PA (%)	94.82 (1.80)	97.85 (1.11)	97.72 (1.04)	89.46 (2.31)	89.55 (3.01)
NA (%)	93.22 (2.72)	97.18 (1.78)	97.86 (1.32)	84.23 (2.75)	83.52 (1.38)
GM (%)	94.00 (1.05)	97.51 (1.21)	97.79 (1.05)	86.79 (1.54)	86.47 (1.37)
OA (%)	93.93 (1.13)	97.50 (1.21)	97.79 (1.05)	86.57 (1.57)	86.21 (1.28)
F1 (%)	93.86 (1.26)	97.49 (1.22)	97.79 (1.05)	86.05 (1.80)	85.63 (1.18)
Time (s)	0.24 (0.05)	0.06 (0.01)	0.07 (0.01)	0.15 (0.18)	0.04 (0.00)
Extracted features	X1–X9	X1, X3, X4, X9	X1, X3	X1, X3, X5, X6, X7	X2, X3, X5, X6, X9

quite close, and are remarkably better than the original set and other feature subsets, and the overall accuracy (OA) reaches 99.31% and 99.40%, respectively. F1 indicators also outperform others. However, the DT#2 feature set only uses 2 features, but can show the same performance. Therefore, the DT#2 feature set is selected as the best feature set of "all features".

The performance evaluation results of "project features" are summarized in Table 11. From Table 11, we can find

that all the metrics of the DT#3 feature set and the DT#4 feature set are quite close, and are dramatically better than the original set and other feature subsets. Their overall accuracies have reached 99.40% and 98.93%, respectively. F1 indicator also points out the same situation. It can be seen that the DT#3 feature set is better than the DT#4 feature set, so the DT#3 feature set is selected as the best feature set of the "project feature".

Table 9 Evaluation results of content features in video game projects

Feature sets Index	Original (11 features)	DT4 (9 features)	DT 5 (5 features)	LASSO3 (5 features)	LASSO4 (3 features)	SVM-RFE3 (5 features)	SVM-RFE4 (1 features)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PA (%)	64.89 (3.65)	64.46 (3.20)	65.63 (5.81)	66.94 (5.50)	64.88 (4.11)	65.06 (5.82)	61.80 (5.34)
NA (%)	62.25 (3.17)	61.96 (3.36)	63.28 (2.22)	62.40 (2.49)	63.75 (4.07)	62.88 (3.32)	58.91 (4.19)
GM (%)	63.54 (3.00)	63.17 (2.68)	64.42 (3.83)	64.61 (3.89)	64.29 (3.58)	63.91 (3.74)	60.34 (4.74)
OA (%)	63.29 (3.02)	62.86 (2.78)	64.21 (3.50)	64.21 (3.70)	64.07 (3.62)	63.50 (3.48)	60.14 (4.69)
F1 (%)	61.19 (4.57)	60.67 (4.75)	63.13 (2.21)	61.64 (2.66)	63.05 (5.23)	62.15 (4.36)	57.04 (5.38)
Time (s)	0.08 (0.03)	0.19 (0.29)	0.19 (0.29)	0.28 (0.51)	0.05 (0.01)	0.64 (1.27)	0.05 (0.01)
Extracted features	X10–X20	X11, X12, X13, X15, X16, X17, X18, X19, X20	X11, X17, X18, X19, X20	X13, X15, X16, X17, X20	X15, X17, X20	X10, X11, X15, X17, X20	X11

Table 10 Evaluation results of all features in mobile game projects

Feature set Index	Original (20 features)	DT#1 (3 features)	DT#2 (2 features)	LASSO#1 (2 features)	SVM-RFE#1 (10 features)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PA (%)	72.54 (5.31)	98.75 (1.28)	99.39 (1.36)	78.95 (9.48)	79.51 (3.73)
NA (%)	94.94 (0.33)	99.37 (0.34)	99.40 (0.33)	94.85 (1.61)	94.26 (0.28)
GM (%)	82.94 (3.11)	99.06 (0.69)	99.39 (0.66)	86.43 (5.65)	86.55 (1.91)
OA (%)	93.55 (0.60)	99.31 (0.35)	99.40 (0.30)	94.01 (1.58)	93.55 (0.13)
F1 (%)	58.10 (3.94)	96.17 (1.97)	96.64 (1.68)	57.74 (15.72)	53.77 (2.19)
Time (s)	0.40 (0.10)	0.06 (0.01)	0.06 (0.01)	0.11 (0.09)	0.26 (0.13)
Extracted features	X1–X20	X1, X3, X9	X1, X3	X1, X3	X3, X5, X7, X9, X10, X11, X14, X15, X16, X19

Table 11 Evaluation results of project features in mobile game projects

Feature set Index	Original (9 features)	DT#3 (2 features)	DT#4 (1 features)	LASSO#2 (1 features)	SVM-RFE#2 (5 features)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PA (%)	90.30 (14.79)	99.39 (1.36)	100.00 (0.00)	82.38 (6.85)	75.73 (11.79)
NA (%)	98.44 (1.21)	99.40 (0.33)	98.84 (0.41)	94.30 (1.62)	94.19 (1.70)
GM (%)	94.00 (8.28)	99.39 (0.66)	99.42 (0.21)	88.07 (3.56)	84.20 (6.48)
OA (%)	97.50 (2.18)	99.40 (0.30)	98.93 (0.38)	93.68 (1.20)	92.92 (1.12)
F1 (%)	86.42 (10.66)	96.64 (1.68)	93.80 (2.28)	52.32 (16.24)	48.59 (15.59)
Time (s)	0.36 (0.11)	0.06 (0.01)	0.04 (0.01)	3.33 (3.89)	0.10 (0.03)
Extracted features	X1–X9	X1, X3	X1	X3	X3, X5, X6, X7, X8

Table 12 Evaluation results of content features in mobile game projects

Feature set Index	Original (11 features) Mean (SD)	LASSO#3 (3 features) Mean (SD)	SVM-RFE#3 (6 features) Mean (SD)	SVM-RFE#4 (3 features) Mean (SD)
PA (%)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
NA (%)	90.77 (0.05)	90.77 (0.05)	90.77 (0.05)	90.77 (0.05)
GM (%)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
OA (%)	90.77 (0.05)	90.77 (0.05)	90.77 (0.05)	90.77 (0.05)
F1 (%)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Time (s)	0.06 (0.00)	0.06 (0.00)	0.06 (0.00)	0.05 (0.01)
Extracted features	X10–X20	X12, X13, X15	X10, X13, X15, X17, X18, X19	X15, X17, X19

Table 12 lists the evaluation results of the "content features". From Table 12, it can be seen that the performance evaluation results of the original feature set and all feature subsets show that PA is 0. It means that the established classifier cannot identify any minority examples (successful projects). This is a typical class imbalance problem. Probably because the number of failed projects is 10 times the number of successful projects, and the attribute values of the content features are not very similar between successful and failed projects. Therefore, the constructed classifier cannot recognize any successful project (positive example).

Due to the class imbalance problems caused by skew class distribution in "content feature" data sets, we will use under-sampling and over-sampling (Synthetic Minority Oversampling Technique, SMOTE) to process the imbalanced data, and re-do feature selection and re-build classifiers. The under-sampling method is to randomly sample from the majority class examples (failed projects), and extract the same number of samples as the minority class examples (successful projects) to rebalance. Then, we use three feature selection methods to select feature candidate subsets, and use SVM for performance evaluation. After the under-sampling method, there are 336 successful projects and 336 failed projects.

SMOTE is a well-known over-sampling method. The main principle is to duplicate minority class examples (successful projects) until the number of samples is equal to the majority class examples (failed projects). After the processing of data oversampling, three feature selection methods are used for feature selection to find feature subsets, and finally SVM is used for performance evaluation. After SMOTE processing, the number of successful projects and failed projects were 3306, respectively.

The results of under-sampling are shown in Table 13. According to Table 13, it can be seen that all performance indicators of all feature subsets are quite close. The under-sampling technique has solved class imbalance problems. Therefore, the SVM–RFE-II feature set with the highest OA is selected as the best feature set for under-sampling.

The results of SMOTE are shown in Table 14. According to this table, it can be seen that all indicators of the original feature set and DT-III feature set are better than the LASSO-II feature set and SVM–RFE-III feature set. But, the original feature set and DT-III feature set have the same features. It's meaningless if we select DT-III. Therefore, only the LASSO-II feature set could be compared with the SVM–RFE-III feature set. All indicators of the LASSO-II feature set are superior to the SVM–RFE-III feature set.

Table 13 Results of implementing under-sampling (content features of mobile game projects)

Feature set Index	Original (11 features) Mean (SD)	DT-I (10 features) Mean (SD)	DT-II 2 features Mean (SD)	LASSO-I 4 features) Mean (SD)	SVM-RFE-I 5 features) Mean (SD)	SVM-RFE-II 3 features) Mean (SD)
PA (%)	59.91 (3.84)	59.57 (3.83)	61.86 (3.55)	64.20 (8.65)	60.89 (1.90)	60.88 (3.03)
NA (%)	59.92 (5.66)	60.17 (5.16)	61.76 (6.77)	55.86 (5.75)	60.69 (4.76)	62.01 (5.40)
GM (%)	59.90 (4.65)	59.86 (4.43)	61.74 (4.29)	59.58 (2.58)	60.77 (3.18)	61.43 (4.12)
OA (%)	59.82 (4.56)	59.81 (4.36)	61.15 (4.19)	61.15 (3.91)	60.56 (3.22)	61.30 (4.02)
F1 (%)	58.99 (6.16)	59.93 (5.29)	59.04 (9.28)	63.95 (7.69)	59.19 (6.86)	61.36 (6.18)
Time (s)	0.04 (0.01)	0.03 (0.01)	0.03 (0.00)	0.03 (0.00)	0.03 (0.00)	0.03 (0.00)
Extracted features	X10–X20	X10, X12, X13, X14, X15, X16, X17, X18, X19, X20	X15, X16	X12, X13, X15, X18	X15, X17, X18, X19, X20	X15, X17, X20

Table 14 Results of implementing SMOTE (content features of mobile game projects)

Feature set	Original (11 features)	DT-III (11 features)	LASSO-II (7 features)	SVM-RFE-III (5 features)
Index	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PA (%)	80.77 (3.75)	80.77 (3.75)	67.81 (4.81)	66.72 (6.39)
NA (%)	79.88 (1.75)	79.88 (1.75)	67.50 (3.62)	63.50 (3.69)
GM (%)	80.32 (2.74)	80.32 (2.74)	67.65 (4.15)	65.08 (4.99)
OA (%)	80.36 (2.91)	80.36 (2.91)	67.62 (4.23)	65.15 (4.95)
F1 (%)	81.58 (2.17)	81.58 (2.17)	70.88 (2.89)	67.94 (2.84)
Time (s)	0.72 (0.03)	0.72 (0.03)	0.93 (0.63)	1.13 (0.70)
Extracted features	X10–X20	X10–X20	X11, X12, X13, X15, X17, X18, X19	X11, X12, X14, X15, X20

Table 15 Comparison of under-sampling and SMOTE in content features

Feature set	Under-sampling SVM-RFE-II (3 features)	SMOTE LASSO-III (7 features)
Index	Mean (SD)	Mean (SD)
PA (%)	60.88 (3.03)	67.81 (4.81)
NA (%)	62.01 (5.40)	67.50 (3.62)
GM (%)	61.43 (4.12)	67.65 (4.15)
OA (%)	61.30 (4.02)	67.62 (4.23)
F1 (%)	61.36 (6.18)	70.88 (2.89)
Time (s)	0.03 (0.00)	0.93 (0.63)
Extracted features	X15, X17, X20	X11, X12, X13, X15, X17, X18, X19

Therefore, the LASSO-II feature set is selected as the best feature set for the over-sampling of "content features".

Table 15 summarizes the comparison results of under-sampling and SMOTE. It can be seen that all the indicators of SMOTE's LASSO-III feature set are better than the under-sampling SVM–RFE-III feature set, so finally the over-sampling LASSO-III feature set is selected as the best feature set.

4.4.3 Concluding remarks

The summary of all selected important features is shown in Table 16. In terms of content features, we have identified three features for the video game project, namely Promotion (X15), Sound effects (X17), and Appearance (X20). For the mobile game project, we also found 7 important features, namely Interaction (X11), Role (X12), System (X13), Promotion (X15), Sound effects (X17), Operation (X18), and Price (X19). Initially, the selected features of the two

Table 16 The discovered important features for video and mobile game crowdfunding projects

Feature	Experiment#1 Video games			Experiment#2 Mobile games		
Index	All	Project	Content	All	Project	Content
	DT 1	DT 3	LASSO4	DT#2	DT#3	SMOTE LASSO-III
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PA (%)	98.02 (1.11)	97.72 (1.04)	64.88 (4.11)	99.39 (1.36)	99.39 (1.36)	67.81 (4.81)
NA (%)	97.37 (2.41)	97.86 (1.32)	63.75 (4.07)	99.40 (0.33)	99.40 (0.33)	67.50 (3.62)
GM (%)	97.68 (0.75)	97.79 (1.05)	64.29 (3.58)	99.39 (0.66)	99.39 (0.66)	67.65 (4.15)
OA (%)	97.64 (0.78)	97.79 (1.05)	64.07 (3.62)	99.40 (0.30)	99.40 (0.30)	67.62 (4.23)
F1 (%)	97.62 (0.83)	97.79 (1.05)	63.05 (5.23)	96.64 (1.68)	96.64 (1.68)	70.88 (2.89)
Time (s)	0.06 (0.02)	0.07 (0.01)	0.05 (0.01)	0.06 (0.01)	0.06 (0.01)	0.93 (0.63)
Important features	Target amount (X1), Update (X3), Number of investors (X9)	Target amount (X1), update (X3)	Promotion (X15), Sound effects (X17), Appearance (X20)	Target amount (X1), Update (X3)	Target amount (X1), Update (X3)	Interaction (X11), Role (X12), System (X13), Promotion (X15), Sound effects (X17), Operation (X18), Price (X19)

games are different, and the emphasis on the narrative content of successful fundraising projects is also quite different. Because video games have physical goods, promotion, sound effects, and appearance will be very important. However, mobile games are mainly downloaded by players, and then through in-game purchases as the main profit model, so the emphasis is on price, interactivity, fluency of operation, and game experiences.

5 Discussions and conclusions

5.1 Discussions

This study uses "video game" and "mobile game" project in Kickstarter.com as our research objects. Through experiments, we find out important factors shown in Table 16. From this table, we can see that the target amount, number of updates, promotion, and sound effects are common important factors for video game and mobile game projects. This study regards them as the key factors affecting the success of game crowdfunding projects.

In both experiements, results of feature section indicated that the target amount and the number of updates are two key factors that affect the success of game crowdfunding. However, judging from the performance of using project features and content features to build a classifier, the classification accuracy of project features is much higher than that of content features. This may be that the amount of

information brought by the content features is less than the project features.

In addition, there are several differences in important factors between video games and mobile games. In all the features, the project features of video games have one additional feature "number of investors", but the number of investors is a factor beyond the control of fundraisers. Therefore, this can only show that the number of investors will affect the success of video game projects. The game is less affected. In terms of content features, in addition to the common "promotion" and "sound effects", "appearance" is also an important success factor for video games. And "interaction", "role", "system", "operation" and "price" are key successful factor for mobile games.

We also analyzed the important factors of the two types of projects as shown in Tables 17 and 18. The analysis results of the video game project are shown in Table 17. In the project features, we can see that the target amount of successful projects below 5000 USD is 46% higher than the 34.1% of failed projects. The proportion of successful projects with more than 5 updates (84.9%) is also much higher than the 18.9% of failed projects. In the content features, the number of mentions of "promotion" and "appearance" related words in the introduction description of successful projects is also significantly more than that of failed projects.

Table 18 summarizes the analysis results of the mobile game project. In the project features, we can see that more than half of the successful fundraising projects (53.6%) have a target amount of less than 5000 USD, while only 33.2% of the failed projects are less than 5000 USD. Moreover, the

Table 17 Analysis of selected important feature between successful and failed video games projects

Project type Important features	Succesful	Failed
Target amount (project)	< 5000 USD46.0% 5000–9999 USD: 14.3% 10,000–19,999 USD: 17.0% 20,000–49,999 USD: 15.3% > 49,999 USD: 7.4%	< 5000 USD: 34.1% 5000–9999 USD: 18.1% 10,000–19,999 USD: 17.7% 20,000–49,999 USD: 14.7% > 49,999 USD: 15.3%
Updates (project)	< 5 times: 15.1% 5–9 times: 20.9% 10–15 times: 18.3% > 15 times: 45.7%	< 5 times: 81.1% 5–9 times: 15.0% 10–15 times: 2.3% > 15 times: 1.6%
Number of investors (project)	< 100 persons: 33.1% 100–499 persons: 33.9% 500–999 persons: 15.1% > 999 persons: 17.9%	< 100 persons: 92.7% 100–499 persons: 6.9% 500–999 persons: 0.4% > 999 persons: 0.0%
Promotion (content)	0–5: 43% 6–10: 35% > 10: 22%	0–5: 67% 6–10: 23% > 10: 10%
Sound effects (content)	0–5: 95% 6–10: 5%	0–5: 98% 6–10: 2%
Appearance (content)	0–5: 66% 6–10: 22% > 10: 12%	0–5: 88% 6–10: 11% > 10: 1%

Table 18 Analysis of selected important feature between successful and failed mobile games projects

Project type Important features	Succeful	Failed
Target amount (project)	<5000 USD: 53.6% 5000–9999 USD: 18.8% 10,000–19,999 USD: 14.0% 20,000–49,999 USD: 9.8% >49,999 USD: 3.9%	<5000 USD: 33.2% 5000–9999 USD: 15.1% 10,000–19,999 USD: 21.1% 20,000–49,999 USD: 15.1% >49,999 USD: 15.5%
Updates (project)	<5 times: 29.8% 5–9 times: 21.7% 10–15 times: 14.6% >15 times: 33.9%	<5 times: 90.8% 5–9 times: 7.3% 10–15 times: 1.2% >15 times: 0.6%
Interaction (content)	0–5: 68% 6–10: 25% >10: 7%	0–5: 83% 6–10: 14% >10: 3%
Role (content)	0–5: 31% 6–10: 35% >10: 34%	0–5: 56% 6–10: 24% >10: 19%
System (content)	0–5: 54% 6–10: 37% >10: 9%	0–5: 75% 6–10: 22% >10: 4%
Promotion (content)	0–5: 55% 6–10: 33% >10: 12%	0–5: 74% 6–10: 20% >10: 6%
Sound effects (content)	0–5: 99% 6–10: 1%	0–5: 99% 6–10: 1%
Operation (content)	0–5: 94% 6–10: 5% >10: 1%	0–5: 96% 6–10: 3% >10: 1%
Price (content)	0–5: 93% 6–10: 7%	0–5: 96% 6–10: 4%

number of updates for successful projects is much higher than that for failed projects. According to the examples we have collected, the number of successful projects has been updated more than five times, accounting for 70.2%, while the number of failed projects has been updated more than five times, accounting for only 9.2%. In the content feature part, we calculate based on the frequency of related words appearing. We can see that interaction, role, system, promotion and other factors are obviously mentioned more frequently in successful fundraising projects than failed projects.

5.2 Conclusion

This study attempts to discover the key successful factors of project descriptions that influence the success of video game and mobile game crowdfunding projects. We proposed a lexicon-based feature selection method which uses real crowdfunding projects on Kickstarter, and define 9 candidate project features that might affect the success of the project from related literature, and summarize the game-related aspects and vocabulary to establish lexicons to form 11 candidate content features. Then, three feature selection

methods including DT, LASSO and SVM–RFE have been employed to select the candidate feature subsets. Finally, SVM is used to evaluate the performance of candidate feature subsets and determine the key succesul factors of game crowdfunding projects.

We consolidate the important factors of the two types of projects as the key factors affecting the success of the game crowdfunding project and give some suggestions to help improve the success rate of crowdfunding projects, as shown in Table 19. From this table, we can see that the target amount, number of updates, promotion (content feature) and sound effects (content feature) are the key factors that affect the success of the game crowdfunding project.

The target amount is a controllable factor for fundraisers. It is recommended to set it below US$5000, and the success rate will be higher. "Update" frequency is also a controllable factor. It is recommended that fundraisers should make more updates during the fundraising period to avoid information asymmetry, so that investors can keep abreast of the situation and interact with investors more. In addition, fundraisers are suggested that when writing a project introduction, more words related to "promotion", "discount", "offer", "music" and "sound" should be used, which will help increase the probability of success of the project.

Table 19 Suggestions of the selected important features

Important features	Definition	Suggestions
Target amount (project)	The fundraising goal set when the project is established	That's the fundraiser's controllable factor. It's recommended to be set within a reasonable range. According to the statistical analysis of the data of this research, the success rate of less than $5000 is higher
Updates (project)	The total number of updates at the end of the project	It's the fundraiser controllable factor. It is recommended that more updates should be made during the fundraising period to avoid information asymmetry and allow investors to keep abreast of the information and current situation
Promotion (content)	Related vocabulary related to "promotion"	It is recommended that fundraisers mention various vocabulary related to "promotion", "discount" and "offer" in the project description to help increase the success rate of fundraising
Sound effects (content)	Related vocabulary related to "sound effects"	It is recommended that fundraisers mention various vocabulary related to "music" and "sound" in the project description, which will help increase the success rate of fundraising

In addition to the above-mentioned key factors, there are some differences between the important features of video games and mobile games. In the project features, the number of investors is one of the important factors of video games. The number of investors representing investors will affect the success of video game projects. But, it's not a controllable factor.

In terms of content features, "appearance" is also one of the important features of video games, which means that when investors consider whether to invest in video game projects, the appearance and packaging are the parts that investors will pay attention to, and more attention should be paid to appearance and packaging. "Interaction", "role", "system", "operation", and "price" are important features of mobile games. It means that investors who invest in mobile games pay more attention to the interactivity, character design, system stability, and operation, control and price.

Moreover, according to results of feature selection, we can find that the importance of the project feature is higher than the content feature. On the one hand, the importance of the project feature is higher, and on the othr hand, the lexicons may not be complete. The lexicon can be improved in the future. Therefore, it is recommended that fundraisers who create game fundraising projects should pay more attention to the setting of project features in addition to the description of the text. For example, the "target amount" should be set appropriately. The project should be updated regularly so that information can be communicated more quickly and transparently, so that investors can better understand the investment project.

For potential direction of future works, other types of game projects or other well-known crowdfunding platforms can be considered in the future. The number of data can be increased in the future to make the research more accurate. Other feature selection methods can be included to find the most suitable feature selection method. Finally, to extend the built lexicons might be considered in the future. Regarding COVID-19, the epidemic has already had a huge impact on the entire world. Future researchers can focus on the change of project types or comments of participants before and during the epidemic to find out the impact of the epidemic on the direction of fundraising.

References

Borrero-Dominguez C, Cordon-Lagares E, Hernandez-Garrido R (2020) Analysis of success factors in crowdfunding projects based on rewards: a way to obtain financing for socially committed projects. Heliyon 6(4):e03744

Bretschneider U, Knaub K, Wieck E (2014) Motivations for crowdfunding: what drives the crowd to invest in start-ups?. In: Twenty second European conference on information systems (ECIS) Tel Aviv, 9–11 June

Caci B, Scrima F, Tabacchi ME, Cardaci M (2018) The reciprocal influences among motivation, personality traits, and game habits for playing Pokémon GO. Int J Hum Comput Interact 35(14):1303–1311

Chan DY-C, Xing Z-L, Zhang Z-H (2017) Antecedents of loyalty in mobile social network games. Adv Soc Sci Educ Hum Res 155:136–140

Chang J-R, Chen M-Y, Chen L-S, Tseng S-C (2019) Why customers don't revisit in tourism and hospitality industry? IEEE Access 7(7):146588–146606

Chang J-R, Chen M-Y, Chen L-S, Chien W-T (2020) Recognizng important factors of influencing trust in O2O models: an example of OpenTable. Soft Comput 24:7907–7923

Chen L-S, Liu C-H, Hsu C-C, Lin C-S (2010) C-Kano model: a new approach for discovering attractive quality elements. Total Qual Manag Bus Excell 21(11):1189–1214

Chen L-S, Lin Z-C, Chang J-R (2015) FIR: an effective scheme for extracting useful metadata from social media. J Med Syst. https://doi.org/10.1007/s10916-015-0333-0

Chen W-K, Riantama D, Chen L-S (2021) Using a text mining approach to hear voices of customers from social media toward the fast-food restaurant industry. Sustainability 13(1):268

Cholakova M, Clarysse B (2015) Does the possibility to make equity investments in crowdfunding project crowd out reward-based investment? Entrep Theory Pract 39(1):145–172

Colgren D (2014) The rise of crowdfunding: social media, big data, cloud technologies. Strat Financ 95(10):56–57

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Cosh A, Cumming D, Hughes A (2009) Outside entrepreneurial capital. Econ J 119(540):1494–1533

Cox J, Nguyen T, Kang SM (2018) The kindness of strangers? An investigation into the interaction of funder motivations in online crowdfunding campaigns. Kyklos 71(2):187–212

Cumming DJ, Hornuf L, Karami M, Schweizer D (2020) Disentangling crowdfunding from fraudfunding. Max Planck Inst Innov Compet Res 16:1–86

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(3):131–156

Devi SG, Sabrigiriraj M (2018) Feature selection, online feature selection techniques for big data classification: a review. In: 2018 International conference on current trends toward converging technologies, p 18637371

Dreisbach C, Koleck TA, Bourne PE, Bakken S (2019) A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 125:37–46

Du Q, Qiao Z, Fan W, Zhou M, Zhang X, Wang AG (2015) Money talks: a predictive model on crowdfunding success using project description. In: Twenty-first Americas conference on information systems, Puerto Rico, pp 1–8

Fernandez-Blanco A, Villanueva-Balsera J, Rodriguez-Montequin V, Moran-Palacios H (2020) Key factors for project crowdfunding success: an empirical study. Sustainability 12(2):599

Frank O, Brors B, Fabarius A, Li L, Haak M, Merk S, Schwindel U, Zheng C, Müller MC, Gretz N, Hehlmann R, Hochhaus A, Seifarth W (2006) Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. Leukemia 20(8):1400–1407

Gamal D, Alfonse M, El-Horbaty E-SM, Salem A-BM (2019) Implementation of machine learning algorithms in Arabic sentiment analysis using N-Gram features. Proced Comput Sci 154:332–340

Gerber EM, Hui JS, Kuo PY (2012) Crowdfunding: why people are motivated to post and fund projects on crowdfunding platforms. In: Proceedings of the international workshop on design, influence, and social technologies: techniques, impacts and ethics

Guenther J, Sawodny O (2019) Feature selection for thermal comfort modeling based on constrained LASSO regression. IFAC Pap OnLine 52(15):400–405

Hollas J (2013) Is crowdfunding now a threat to traditional finance? Corp Financ Rev 2013:27–31

Hsiao K-L, Chen C-C (2016) What drives in-app purchase intention for mobile games? An examination of perceived values and loyalty. Electron Commer Res Appl 16:18–29

Kim T, Por MH, Yang S-B (2017) Winning the crowd in online fundraising platforms: the roles of founder and project features. Electron Commer Res Appl 25:86–94

Krammer P, Habala O, Mojzis J, Hluchy L, Jurkovic M (2019) Anomaly detection method for online discussion. Proced Comput Sci 155:311–318

Kwon TH, Zaman M, Song S, Kim Y (2020) Profiling thrift store shoppers with a decision tree predictive model. Sustain Fash. https://doi.org/10.31274/susfashion.11459

Lagazio C, Querci F (2018) Exploring the multi-sided nature of crowdfunding campaign success. J Bus Res 90:318–324

Lax A (2017) Success factors of game industry crowdfunding campaigns. Master's thesis, Faculty of Information Technology, University of Jyväskylä, pp 1–63

Li H, Chen X, Zhang Y, Hai M, Hu H (2018) A new factor in the effect of crowdfunding projects. Proced Comput Sci 139:114–119

Liang X, Hu X, Jiang J (2020) Research on the effects of information description on crowdfunding success within a sustainable economy—the perspective of information communication. Sustainability 12(2):650

Liu W, Zhai J, Ding H, He X (2017) The research of algorithm for protein subcellular localization prediction based on SVM-RFE. In: 2017 10th International congress on image and signal processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, pp 1–6

Loureiro SMC, Guerreiro J, Ali F (2020) 20 years of research on virtual reality and augmented reality in tourism context: a text-mining approach. Tour Manag 77:104028

Ma L, Destercke S, Wang Y (2016) Online active learning of decision trees with evidential data. Pattern Recogn 52:33–45

McEligot AJ, Poynor V, Sharma R, Panangadan A (2020) Logistic LASSO regression for dietary intakes and breast cancer. Nutrients 12(9):2652

Meyskens M, Bird L (2015) Crowdfunding and value creation. Entrep Res J 5(2):155–166

Miner G, Dursun D, John E, Andrew F, Thomas H, Robert AN (2012) Practical text mining and statistical analysis for non-structured text data applications. Elsevier, Waltham

Mollick E (2014) The dynamics of crowdfunding: an exploratory study. J Bus Ventur 29(1):1–16

Mollick E, Robb A (2016) Democratizing innovation and capital access: the role of crowdfunding. Calif Manag Rev 58(2):72–87

Namazkhan M, Albers C, Steg L (2020) A decision tree method for explaining household gas consumption: the role of building characteristics, socio-demographic variables, psychological factors and household behavior. Renew Sustain Energy Rev 119:109542

Paul S, Magdon-Ismail M, Drineas P (2016) Feature selection for linear SVM with provable guarantees. Pattern Recogn 60:205–214

Petruzzelli AM, Natalicchio A, Panniello U, Roma P (2019) Understanding the crowdfunding phenomenon and its implications for sustainability. Technol Forecast Soc Chang 141:138–148

Ramadani V (2009) Business angels: who they really are. Strat Change 18(7–8):249–258

Schmidt E, Gatsis N, Akopian D (2020) A GPS spoofing detection and classification correlator-based technique using the LASSO. IEEE Trans Aerosp Electron Syst. https://doi.org/10.1109/TAES.2020.2990149

Schuckert M, Liu X, Law R (2015) A Segmentation of online reviews by language groups: how English and non-English speakers rate hotels differently. Int J Hosp Manag 48:143–149

Schwienbacher A, Larralde B (2010) Crowdfunding of small entrepreneurial ventures. SSRN Electron J 10:1–23

Shao Z, Yang S, Gao F, Zhou K, Lin P (2017) A new electricity price prediction strategy using mutual information-based SVM-RFE classification. Renew Sustain Energy Rev 70:330–341

Shneor R, Munim ZH (2019) Reward crowdfunding contribution as planned behaviour: an extended framework. J Bus Res 103:56–70

Short JC, Ketchen DJ, Mckenny AF, Allison TH, Ireland RD (2017) Research on crowdfunding: reviewing the (very recent) past and celebrating the present. Entrep Theory Pract 41(2):149–160

Smith AN (2015) The backer–developer connection: exploring crowdfunding's influence on video game production. New Media Soc 17(2):198–214

Sok HK, Ooi MPL, Kuang YC, Demidenko S (2016) Multivariate alternating decision trees. Pattern Recogn 50:195–209

Thomaz GM, Biz AA, Bettoni EM, Mendes-Filho L, Buhalis D (2017) Content mining framework in social media: a FIFA World Cup 2014 case analysis. Inf Manag 54(6):786–801

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc. Ser B (Method) 58(1):267-288

Vanacker TR, Manigart S (2010) Pecking order and debt capacity considerations for high-growth companies seeking financing. Small Bus Econ 35(1):53–69

Vijayakumar B, Muhammad Fuad MM (2019) A new method to identify short-text authors using combinations of machine learning and natural language processing techniques. Proced Comput Sci 159:428–436

Wang W, Zhu K, Wang H, Wu Y-CJ (2017) The impact of sentiment orientations on successful crowdfunding campaigns through text analytics. IET Softw 11(5):229–238

Wang S, Ji B, Zhao J, Liu W, Xu T (2018) Predicting ship fuel consumption based on LASSO regression. Transp Res D: Transp Environ 65:817-824

Wang C, Xiao Z, Wang B, Wu J (2019) Identification of autism based on SVM-RFE and stacked sparse auto-encoder. IEEE Access 7:118030–118036

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. In: A volume in The Morgan Kaufmann series in data management systems

Xu B, Zheng H, Xu Y, Wang T (2016) Configurational paths to sponsor satisfaction in crowdfunding. J Bus Res 69(2):915–927

Yang FJ (2019) An extended idea about decision trees. In: 2019 International conference on computational science and computational intelligence. https://doi.org/10.1109/CSCI49370.2019.00068

Yang Y, Bi C, Liu L (2020) Profit allocation in investment-based crowdfunding with investors of dynamic entry times. Eur J Oper Res 280(1):323–337

Zhang H, Chen W (2019) Crowdfunding technological innovations: interaction between consumer benefits and rewards. Technovation 84:11–20

Zhao F, Fang X (2009) Factors affecting online game players' loyalty. In: International conference on internationalization, design and global development, pp 197–206

Zhao J, Zhang C, Yao F (2020) Efficient selection of visual features in automatic image retrieval. In: 2020 12th International conference on measuring technology and mechatronics automation, pp 361–365

Zhong B, Pan X, Love PED, Sun J, Tao C (2020) Hazard analysis: a deep learning and text mining framework for accident prevention. Adv Eng Inform 46:101152

Zhou MJ, Lu B, Fan WP, Wang GA (2018) Project description and crowdfunding success: an exploratory study. Inf Syst Front 20(2):259–274

Ziegler T, Shneor R, Garvey K, Wenzlaff K, Yerolemou N, Zhang B, Hao R (2018) Expanding horizons: the 3rd European alternative finance industry report. Cambridge Centre for Alternative Finance, UK