Decision Support Systems 91 (2016) 67–76

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

The determinants of crowdfunding success: A semantic text analytics approach

Hui Yuan [a], Raymond Y.K. Lau [a], Wei Xu [b,c,*]

[a] *Department of Information Systems, College of Business, City University of, Hong Kong, Hong Kong Special Administrative Region*
[b] *School of Information, Renmin University of China, Beijing 100872, PR China*
[c] *Smart City Research Center, Renmin University of China, Beijing 100872, PR China*

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 1 August 2016
Accepted 1 August 2016
Available online 6 August 2016

Keywords:
Crowdfunding
Topic modeling
Text analytics
Machine learning

ABSTRACT

In the era of the Social Web, crowdfunding has become an increasingly more important channel for entrepreneurs to raise funds from the crowd to support their startup projects. Previous studies examined various factors such as project goals, project durations, and categories of projects that might influence the outcomes of the fund raising campaigns. However, textual information of projects has rarely been studied for analyzing crowdfunding successes. The main contribution of our research work is the design of a novel text analytics-based framework that can extract latent semantics from the textual descriptions of projects to predict the fund raising outcomes of these projects. More specifically, we develop the Domain-Constraint Latent Dirichlet Allocation (DC-LDA) topic model for effective extraction of topical features from texts. Based on two real-world crowdfunding datasets, our experimental results reveal that the proposed framework outperforms a classical LDA-based method in predicting fund raising success by an average of 11% in terms of F_1 score. The managerial implication of our research is that entrepreneurs can apply the proposed methodology to identify the most influential topical features embedded in project descriptions, and hence to better promote their projects and improving the chance of raising sufficient funds for their projects.

1. Introduction

In the era of the Social Web, crowdfunding has become an increasingly more important way for entrepreneurs or small enterprises to raise the essential capitals from the crowd to support their projects or businesses. Crowdfunding websites such as Kickstarter and IndieGoGo behave as online intermediary agents that allow projects founders to quickly reach a large number of individual investors with minimal costs. It is believed that crowdfunding helps converting ordinary customers to business investors [1]. Not only does the idea of crowdfunding emerge in developed countries but it also becomes very popular in developing countries such as China. According to the statistics revealed at the crowdfunding website named Zhongchou[1] in China, there were 15,073 projects supported by 802,308 backers who contributed a total of $171,753,514 (RMB) by 2014. Dreamore, Demohour, and Zhongchou are among the most popular crowdfunding websites in China. However, the percentage of crowdfunding projects that can reach their original fund raising goals is relatively small among all crowdfunding websites. For example, only 44% of the projects can reach their initial fund raising goals at Kickstarter [2]. For most crowdfunding websites, if a project cannot raise sufficient fund in time with respect to the original funding goal, the project will be marked as a failure and its fund raising campaign will stop. This becomes the main hindrance of leveraging crowdfunding to acquire venture capitals for supporting a variety of innovative business ideas. In this paper, we focus on examining the influential features, particularly textual features which may affect fund raising successes of crowdfunding projects. For brevity, crowdfunding success refers to the fund raising success of a project for the rest of this paper.

Previous research has examined the dynamics of founders and backers on crowdfunding platforms [1,3,4,5,6,7]. However, these studies mainly examined crowdfunding platforms of the developed countries such as Kickstarter and IndieGoGo that used English as the main communication language. Few studies have investigated into the crowdfunding dynamics in developing countries such as China. Moreover, though some studies explored the impact of numerical features (e.g., funding amount, duration of project, etc.) on crowdfunding success [1,8,9], none of the previous study examined the influence of topical features (i.e., latent semantics) mined from textual descriptions of projects on fund raising success. Our research work is just able to fill the aforementioned research gaps.

For crowdfunding, project founders raise funds by describing their projects and offering rewards to investors (i.e., backers) via crowdfunding websites. Accordingly, the descriptions of projects and rewards

* Corresponding author at: School of Information, Renmin University of China, Beijing 100872, PR China.
E-mail addresses: hui.yuan@my.cityu.edu.hk (H. Yuan), raylau@cityu.edu.hk (R.Y.K. Lau), weixu@ruc.edu.cn (W. Xu).

[1] http://www.zhongchou.com.

http://dx.doi.org/10.1016/j.dss.2016.08.001

can be leveraged to analyze and predict project success. For instance, we successfully extracted a set of semantically related terms (i.e., a topic) such as {protection, rubbish, environment, ecology, …} by applying a topic modeling method [10] to the project descriptions published at Dreamore. This topical feature turns out to be effective to predict fund raising success because the trend in China and the rest of the world is toward environmental protection, and individuals generally like to support "greeny" projects. For topic modeling, each topic captures a semantically coherent "concept" about the real-world (e.g., environmental protection) [11,12]. One novelty of our research is to combine topical features with common numerical features to enhance the prediction of crowdfunding success. A topic-based text analytics method is different from the traditional keyword-based approach in that a topic consists of a set of semantically coherent words, whereas the keyword-based method assumes the independence among words [11,13]. For instance, given the keyword "apple", it may refer to projects of Apple Inc. that are generally supported by the crowd, or projects related to "apple fruit" that are not necessarily supported by the crowd. In fact, previous research shows that keyword-based method is not as effective as topic-based method in text analytics [11,13,14].

In sum, the main contributions of our research work are threefold. First, we design a novel text analytics framework for analyzing and predicting fund raising successes of crowdfunding projects. Second, we develop the domain-constraint LDA (DC-LDA) topic model for more effective mining of topical features (i.e., latent semantics) from textual descriptions of projects. Finally, we performed an empirical analysis to identify the discriminatory features that influence fund raising successes of projects based on real-world crowdfunding platforms. To the best of our knowledge, this is the first successful research of applying a topic modeling method to extract topical features from project descriptions to analyze and predict crowdfunding project success. The managerial implication of our research is that entrepreneurs or small businesses can apply the proposed framework to identify the most influential textual features that affect fund raising outcomes, and hence they can publicize and promote their startup projects by using these features to improve the chance of fund raising success.

The rest of the paper is organized as follows. Section 2 summarizes previous studies related to crowdfunding, and compares our work with the previous studies. Then, the proposed text analytics-based methodology for analyzing and predicting crowdfunding success is highlighted in Section 3. Section 4 illustrates the computational details of the proposed methodology. Discussions of the experimental procedures and the empirical results are given in Section 5. Finally, we offer concluding remarks and pinpoint future directions of our research work.

2. Related work

2.1. An overview of crowdfunding

While crowdsourcing aims to help people perform various tasks by leveraging the "crowd" [15], crowdfunding refers to the completion of a specific type of task, that is, fund raising by using the crowd. Crowdfunding is "an open call, essentially through the Internet, for the provision of financial resources either in form of donation or in exchange for some form of reward and/or voting rights in order to support initiatives for specific purposes" [16]. Crowdfunding consists of three important components: project founders, crowd funders (i.e., backers), and crowdfunding platforms which connect founders to funders. Crowdfunding projects can be classified into different categories such as non-profit or for-profit projects [17]. Alternatively, crowdfunding projects can be categorized as equity-based, reward-based, loan-based, or donation-based projects [18]. Besides, there are two types of founders, namely founders who want to implement their ideas and founders who want to promote their businesses [19]. Previous studies identified the unique advantages of crowdfunding [18,20,21,22,23,24]. Some researchers focused on studying crowdfunding activities for specific business sectors such as the recording industry or the journalism profession [17,25,26,27]. Previous studies also examined the critical success factors of crowdfunding. For example, previous study found that project success rate was improved if a project could successfully acquire most of the required fund [28]. Furthermore, culture and geography are the determinants of successful crowdfunding activities [29].

2.2. The dynamics of crowdfunding

Previous research has also examined the dynamics of crowdfunding. For example, the main motivators for individuals to participate into crowdfunding were to explore innovative ideas, enhance social participation, and obtain financial rewards [1]. The main benefits for founders to engage in crowdfunding were to raise capital, form relationship with backers, obtain approval, maintain control, learn funding skill, and expand their awareness [3]. On the other hand, backers' participations were due to rewards, supporting innovative ideas, and contributing to a community [3]. Employees tended to engage in crowdfunding because of their needs, minority disciplines, technical interests, and constraint removal [4]. The languages used by founders were also explored, and the empirical results showed that reciprocity and scarcity were beneficial to project success [5]. In addition, word-of-mouth (WOM) and observational learning could influence the recipient conversion [6]. The dynamics of crowdfunding were also studied in terms of the phases used in projects and the crowdfunding social networks [7].

2.3. Predicting crowdfunding success

Features such as project goals, project categories, number of rewards, project duration, whether a project was connected to Twitter or Facebook, whether video was present, number of friends in a social network, grade level, and number of sentences in project descriptions were fed to machine learning classifiers such as Support Vector Machine (SVM) and Decision Trees (DT) to predict crowdfunding success [8]. In addition, linguistic features extracted from project descriptions were combined with 59 common features to predict crowdfunding success [9]. Project success was also analyzed in terms of the quality of projects and the sizes of founder networks [5]. Besides, temporal features and machine learning methods such as K-nearest Neighbor (KNN), Markov Chain, and SVM were applied to predict project success [2]. Social feature analysis was also adopted to predict the number of backers and fund raising success [7].

2.4. The main differences between our work and previous studies

Our work differs from the previous studies in three ways. First, while previous studies mainly examined reward-based projects, we aim to study both reward and non-reward based projects. Second, though machine learning methods were applied to predict crowdfunding success, none of the previous work explored topic modeling methods for mining topical features from project descriptions and reward descriptions to predict fund raising success. Third, most of the previous studies were conducted based on crowdfunding platforms that used English as the communication language. Our research aims to analyze crowdfunding success based on Chinese crowdfunding platforms.

3. A text analytics framework for crowdfunding analysis

Though previous studies found that the language used in project descriptions might influence fund raising success [9], only shallow linguistic features (e.g., number of words, spelling errors, etc.) were examined. For the proposed text analytics framework, we advocate a topic modeling method to extract topical features (i.e., latent semantics) from project descriptions and reward descriptions to predict crowdfunding success. In particular, we design a new DC-LDA topic model by extending the classical LDA model [10] to improve the effectiveness of topic mining. The mined topical features together with some common project attributes are then fed into the random forest classifier [30] to predict the fund raising success of a project. The proposed text analytics framework for crowdfunding analysis is outlined in Fig. 1. The proposed framework consists of three main processes, namely data collection, feature extraction, and prediction.

3.1. Data collection

The domain-specific news articles are collected from popular news websites by using the search keyword “crowdfunding”. The news articles are regarded as the prior domain knowledge which is utilized to build a semantic network comprising must-links and cannot-links among words. Such a textual corpus is also applied to perform guided topic mining by using the proposed DC-LDA topic model. The computational details of the DC-LDA model will be illustrated in the following section. Meanwhile, project data from different crowdfunding websites are crawled. Only the expired projects with a confirmed project status (e.g., success or failure) are crawled. The retrieved project data include both numerical features (e.g., durations of projects) and textual features (e.g., project descriptions). As for numerical features, five common attributes as shown in Table 1 are utilized by the proposed prediction model. As for textual features, both project descriptions and reward descriptions are used because the semantics of these descriptions are fundamentally different. For example, project descriptions capture the inherent nature of crowdfunding projects, whereas reward descriptions reveal the various incentives offered to backers.

3.2. Feature extraction

Feature extraction is one of the most important tasks for any classification exercises [31,32,33]. After textual data of projects and relevant online news articles are collected, a word segmentation process is applied to divide each Chinese sentence into words and identify the part-of-speech (POS) of each word. For our study, only nouns are used because they are considered the most representative tokens of a document [34,35]. Afterwards, stop words are removed from sentences. The preprocessed news articles are applied to construct a semantic network of words, which includes both the must-links and the cannot-links relations of words [13,36]. Such a semantic network captures the prior domain knowledge of crowdfunding and it is applied to more effectively mine topical features from project descriptions and reward descriptions. The extracted topical features which represent the latent semantics of the inherent nature of projects and various kinds of rewards are then combined with common numerical features to predict fund raising successes of crowdfunding projects.

Table 1
Numerical project features.

Feature	Description
Goal	Targeted funding amount
Score	The credit score of the creator
Max	Maximum fund to be raised
Min	Minimum fund to be raised
Level	Number of different amounts to be donated

3.3. Prediction

A Random Forest (RF) classifier [30] is proposed to carry out the prediction tasks. RF is a widely used ensemble learning method and it consists of a set of decision trees. Although each constituent decision tree may only be a weak learner, the ensemble of these decision trees usually outperforms well-known classifiers such as SVM for a variety of classification tasks [30,37,38,39]. For our crowdfunding evaluation dataset, each project (i.e., a classification instance) is marked with the class label “1” if the targeted funding goal is achieved; otherwise, it is marked with the class label “-1”. These labels represent the

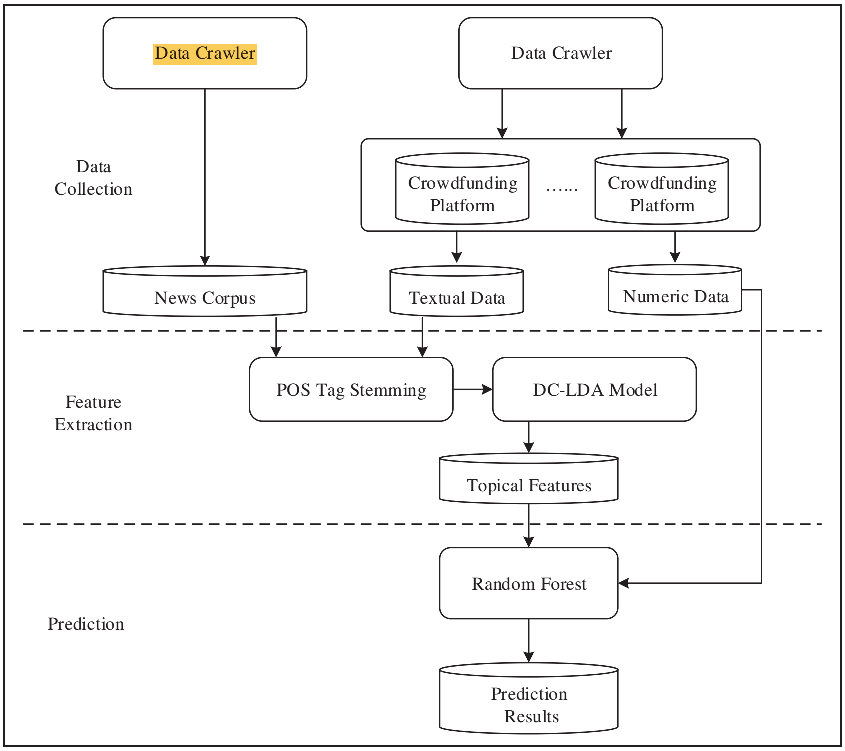


Fig. 1. A text analytics framework for crowdfunding analysis.

ground-truths of the prediction tasks. For our evaluation dataset, a balanced class distribution (i.e., similar number of successful and unsuccessful cases) is maintained because an imbalanced training set may lead to serious learning biases [40,41]. The whole evaluation dataset is divided into a training set and a test set. The training set is applied to train the RF classifier in advance. In particular, each decision tree is trained separately by using a random subset of the entire training set. After the training process, a RF-based predictive model is established to predict if a project from the test set is successful or not. The model's predictions are then compared with the pre-established ground-truths to assess the effectiveness of the proposed framework.

4. The computational methods

4.1. The classical LDA model

Topic modeling methods have been widely used in text mining tasks. Latent Dirichlet Allocation (LDA) is one of the topic modeling methods that introduces a latent variable (i.e., a topic) between the traditional constructs of documents and words to better describe the generation of each document of a textual corpus [10]. The topics represent the latent semantic concepts embedded in documents [11,13]. More specifically, the LDA model utilizes two Dirichlet distributions, namely topic-word distribution and document-topic distribution to describe the generation of words in documents. Essentially, each document is considered as a mixture of some topics and each topic is seen comprising a mixture of words [10].

4.2. The basic intuitions of the proposed DC-LDA model

Although LDA has been widely used in topic modeling, there are some weaknesses of this method [11,13,42]. First, since LDA is an unsupervised learning method, the resulting topics could be very noisy (i.e., some topics contain irrelevant words). Second, as topics may contain irrelevant words, a topic may not be able to represent a semantically coherent concept of a real-world domain; this may lead to the human interpretation problem of mined topics [13]. Third, the random assignment of words to topics may lead to the convergence problem of the iterative topic modeling process [11]. Previous studies examined various methods to alleviate the problems of LDA. For instance, topic-in-set knowledge for guiding the topic modeling process was explored [42]. The must-link (i.e., words sharing similar semantics) and the cannot-link (i.e., words representing quite different meanings) relations were constructed based on an existing corpus, and then applied to guide the LDA model [36]. Moreover, general knowledge is applied to extract prior domain knowledge for guiding topic modeling process [43,44].

To address the shortcomings of the LDA model, we design the DC-LDA model for more effective topic mining. The basic intuition of the DC-LDA model is that we make use of the inherent semantic relations among words of a problem domain to guide the topic modeling process. For instance, since the word "project" and "innovative" are semantically related and they often co-occur in the crowdfunding problem domain, these words are captured via the must-link relation of a domain-specific semantic network. In contrast, "warrant" and "sculpture" are not semantically related under the crowdfunding context because "warrant" is about a stock investment product and "sculpture" is about a totally different type of project of arts. Accordingly, these words are captured via the cannot-link relation of the domain-specific semantic network. By using the must-link and cannot-link relations encoded in the domain-specific semantic network, we can better guide the assignment of words to topics during the iterative topic modeling process. For example, for the topic of "crowdfunding project", the word "innovative" has a better chance to be assigned to the topic, whereas the word "sculpture" is less likely to be assigned to the topic of "financial investment project" according to the cannot-link captured in the semantic network. Due to this guided topic modeling process, the resulting topics are more semantically coherent and easier to be understood by humans. Moreover, the convergence of the iterative topic modeling process is quicker due to the guided topic-word assignments instead of random assignments.

To construct the semantic network, online news articles related to crowdfunding are first retrieved to build a domain-specific corpus. Then, the co-occurrence statistics of each pair of words are collected from this domain-specific corpus. These co-occurrence statistics are applied to estimate the prior probabilities of the must-link and the cannot-link relations among words. At this stage, a domain-specific semantic network is constructed based on the prior probability distributions among each pair of words. Finally, the domain-specific semantic network is applied to the proposed DC-LDA topic model to guide the assignment of words to topics. In other words, domain-specific knowledge (i.e., the prior word probability distributions) is incorporated into the proposed topic modeling method to improve the quality of the mined topics and the speed of convergence.

4.3. The computational details of the DC-LDA model

Fig. 2 shows the proposed DC-LDA model by using the plate notation. Let $\{w_1, w_2, \dots, w_V\}$ denotes the set of vocabulary of a corpus and V is the number of unique words of the corpus. Let K represent the number of topics of the corpus. $D = \{d_1, d_2, \dots, d_M\}$ denotes the set of all documents of the corpus with the corresponding number of words $\{N_1, N_2, \dots, N_M\}$ in each document. α is a Dirichlet prior for establishing the document-topic distribution θ, and β is the Dirichlet prior for estimating the topic-word distribution φ. Different from

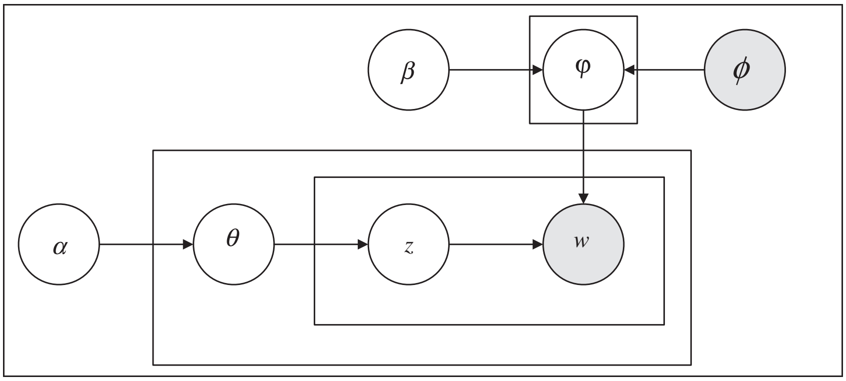


Fig. 2. The DC-LDA Model.

the classical LDA model where the topic-word distribution φ only depends on the prior β, a domain-specific word distribution prior ϕ is applied to establish φ in the proposed DC-LDA model. The domain-specific word distribution prior ϕ is estimated based on the aforementioned semantic network.

The document generation process via the DC-LDA model is described as follows.

Step 1. Choose a topic distribution $\theta_d \sim Dir(\alpha)$.

Step 2. For each topic $k \in \{1, 2, \dots, K\}$, choose a word distribution $\varphi_k \sim Dir(\beta)$.

Step 3. Adjust the topic-word distribution by $\varphi_{k,v} = \varphi_{k,v} \cdot \phi_{k.v}^T$.

Step 4. For each document $d \in \{d_1, d_2, \dots, d_M\}$, the topic distribution for the document is decided. Then, each word is generated through the following two steps.

- Generate one specific topic $z_i \sim Mul(\theta_d)$
- Generate one specific word $w_i \sim Mul(\varphi_{z_i})$

These two steps are iterative until all the words in each document d are generated. To estimate the joint distribution of all the variables including observed and latent variables of the DC-LDA model, Gibbs sampling [45] is applied according to the following procedure.

Step 1: Traverse all the words in the corpus, and then allocate a topic to each word randomly.

Step 2: Add 1 to the four variables $n_m^{(k)}$, n_m, $n_k^{(t)}$,n_k, which represent the number of times topic k is assigned to a document m, the total number of topics assigned to the document m, the number of times word t is assigned to topic k, and the total number of words for topic k.

Step 3: Traverse the corpus again. Assume the word i in document m belongs to topic k, subtract 1 from the four variables and a new topic is allocated to word t according to the following formulation:

$$P(z_i = k | \boldsymbol{z}_{\neg i}, \boldsymbol{w}) \propto \phi_{k,t}^T \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} \left(n_{m,\neg i}^{(t)} + \alpha_k \right)} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} \left(n_{k,\neg i}^{(t)} + \beta_t \right)} \tag{1}$$

where $\phi(V \times K)$ is employed to adjust the generation probability.

Step 4: Steps 2 and 3 are repeated until θ and φ converge, or the maximum number of iteration is reached.

A domain-specific word distribution prior ϕ is required for the DC-LDA model, and this prior is established according to a semantic network *dc* which is constructed based on an online news corpus related to crowdfunding. More specifically, the semantic network is constructed according to the following procedure:

Step 1. A text preprocessing procedure is invoked to extract nouns from the online news corpus.

Step 2. Word co-occurrence statistics are collected from the online news corpus to estimate the initial cluster-word distribution ϑ.

Step 3. All the online news articles are regrouped according to the initial cluster-word distribution ϑ established in Step 2.

Step 4. Step 2 and Step 3 are repeated R number of time to obtain a series of cluster-word distribution ϑ.

For instance, assume that there are 100 initial word clusters and the top 1000 words with the highest probabilities in each cluster are selected. Then, each sentence in the online news corpus is labeled according to the established word cluster. If a sentence contains words of a particular cluster, the sentence will be labeled with that cluster number. If a sentence contains words from multiple clusters, the cluster with the highest word probabilities is assigned to this sentence. After the labeling process, all the sentences with the same cluster number are grouped into one document. Based on the regrouped documents, a sampling process is applied to estimate the initial ϑ. The initial ϑ is then applied to label sentences and regroup documents again. Such a process is repeated until the maximum number of iteration R is reached. So, there are R different cluster-word distributions ϑ at the end. Then, a semantic network, which contains the must-link and the cannot-link relations, is constructed according to the series of cluster-word distributions ϑ. The basic intuition is that if two words are often put under the same cluster, they may have inherent semantic relationship, and so they are captured by the must-link relation of the semantic network *dc*. On the contrary, if two words tend to be allocated to different clusters, it reveals their semantic independence. So, they are encoded in the semantic network by using the cannot-link relation. The above learning process of prior domain knowledge is based on a series of cluster-word distributions $\{\vartheta^1, \vartheta^2, \dots, \vartheta^R\}$. Our semantic network construction process is illustrated in Algorithm 1.

Algorithm 1. Semantic network construction.

```
Inputs: a series of cluster-word distributions$\{\vartheta^1, \vartheta^2, \dots, \vartheta^R\}$, and$\vartheta_{ij}^r$is the probability of ith word in jth cluster at the rth iteration.
Vocabulary set $V = \{w_1, w_2, \dots, w_v\}$, $w_i$ is the ith word, v is the total number of words.
$R$ is the iterative number.
$P_1$ is the number of words with highest probabilities in each cluster for must-link.
$P_2$ is the number of words with highest probabilities in each cluster for cannot-link.
Output: A semantic network dc
  initialize $r = 0$
  while $r$ != R do
      for each row $i$ in $\vartheta^r$
          sort($\vartheta_i^r$)
          add top $P_1$ words =>$ML$
          add top $P_2$ words =>$CL$
      end
      $r$++
  end
must$(m \times m) = 0, m = ML.size$
cannot$(n \times n) = 0, n = CL.size$
while $r$ != R do
      for each row $i$ in $\vartheta^r$
          sort($\vartheta_i^r$)
          add top $P_1$ words =>$A_i^r$
          add top $P_2$ words =>$B_i^r$
          if two words $w_m$, $w_n$ in $A_i^r$
              $must(w_m, w_n)$++ and $must(w_n, w_m)$++
          if two words $w_m$, $w_n$ in $B_i^r$
              $cannot(w_m, w_n)$++ and $cannot(w_n, w_m)$++
      end
      $r$++
end
for each $(w_m, w_n)$
  if $must(w_m, w_n) == R$
      add $(w_m, w_n)$tomust-link
  if $cannot(w_m, w_n) == 0$
      add $(w_m, w_n)$tocannot-link
end
add must-link and cannot-link to semantic network dc
```

After the construction of the semantic network *dc*, we can estimate ϕ_t. Given a word t in the ith position of the mth document and a topic k, browse the semantic network *dc* to verify whether the word t has any relationship with other words (e.g., must-link and cannot-link). If the word t is not related to other words in the semantic network, ϕ_t is set to 1 for all the topics. Otherwise, ϕ_t is estimated based on the number of must-links and the number of cannot-links the word t has. The detailed procedure of estimating ϕ_t is given in Algorithm 2. For instance, *mcount* indicates the number of must-link relations of the word t and *ncount* returns the number of cannot-link relations of t. The termτrefers to the weight of a must-link relation in *dc* and $(1-\tau)$ represents the weight of a cannot-link relation, respectively. In addition, the confidence degree parameter ρ is applied to characterize the reliability of the semantic network that is constructed based on a corpus of online news articles related to crowdfunding. The parameter $\rho$is set within the unit interval [0, 1] with a larger value indicating a higher reliability of the semantic network.

Algorithm 2. Estimating $\phi_{t.}$

```
Inputs: Semantic network dc, including must-link relations dc_M and cannot-link relations dc_C; the current word t, the ith word of mth document; K is the number of topics.
Output: $\phi_t$
Main Procedure:
initialize $mcount_i$, $ncount_i$ and $tcount_i$ with 0
if t in dc
    for each topic k in topics K
        for each word w in topic k
            if (w, t) in dc_M
                $mcount_k$++;
            end if
            if (w, t) in dc_N
                $ncount_k$++;
            end if
        end for
        $tcount_k = mcount_k \cdot \tau - ncount_k \cdot (1-\tau)$ ;
    end for
else
    $tcount = \{1,1,\ldots,1\}$
end if
normalize $tcount$;
$\phi_t = tcount * \rho + (1-\rho)$;
```

As a whole, the proposed DC-LDA model aims to uncover high quality topics that characterize project descriptions (i.e., documents) for predicting fund raising successes of crowdfunding projects. However, we need to specify the number of topics that is utilized by the DC-LDA model to carry out the topic modeling process. We apply the perplexity-based approach [49] to estimate the number of topics applicable to a corpus. Formally, the *Perplexity* measure is defined as follows:

$$perp(D) = exp\left(-\frac{\sum_{d\in D} lnP(d|\theta,\varphi)}{\sum_{d\in D}|d|}\right) \tag{2}$$

where D represents a corpus, d is a document of the corpus D, and $P(d|\theta,\varphi)$ is computed as follows.

$$P(d|\theta,\varphi) = \prod_{w_k\in d}\sum_{z_i\in Z_t} P(w_k|z_i)\cdot P(z_i|d) \tag{3}$$

Once a reasonable number of topics of a corpus are identified, the corresponding topic distribution is obtained by invoking the proposed DC-LDA model. Then, the topic distribution is applied to build topical features for enhancing the prediction process.

5. Experiments and results

5.1. The dataset

Our crowdfunding dataset was retrieved from two popular crowdfunding websites in China, namely Dreamore[2] and Zhongchou by using our dedicated crawlers developed with Python. Moreover, we accessed to the Sina online news forum[3] which is one of the most popular portals in China. All the online news articles referring to "crowdfunding" were crawled for the construction of the semantic network. At the end, 23,113 online news articles were collected from Sina for constructing the semantic network. For the Dreamore website, a total of 500 completed projects were crawled in April 2014, including 250 successful projects and 250 unsuccessful projects of various project categories. For the Zhongchou website, it only shows the successful projects and projects in progress. To identify the unsuccessful projects, we selected the projects in progress based on their deadlines and their secured funds so far. If a project is approaching the deadline and the fund secured is far below the funding goal, the project is likely to fail. We selected a total of 500 projects with 250 unsuccessful projects and 250 successful projects from various project categories of Zhongchou in March 2016. Since the Zhongchou platform does not release the credit score of a founder, only four numerical features such as goal, max, min, and level were utilized. Our evaluation dataset has a balanced class distribution because the same number of successful and unsuccessful cases is included. Finally, our evaluation dataset was divided into a training set of 800 projects and a test set of 200 projects in which a balanced class distribution was maintained.

5.2. The performance measures

To evaluate the effectiveness of the proposed text analytics framework, we applied four common performance measures. Table 2 shows the confusion matrix of the prediction tasks. Based on the confusion matrix, four common performance measures are defined as follows [46].

$$\text{Accuracy}(\%) = \frac{TP+TN}{TP+FP+FN+TN}\times 100 \tag{4}$$

$$\text{Precision}(\%) = \frac{TP}{TP+FP}\times 100 \tag{5}$$

$$\text{Recall}(\%) = \frac{TP}{TP+FN}\times 100 \tag{6}$$

$$F_1(\%) = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}\times 100 \tag{7}$$

5.3. Model parameter estimation

The procedure of cluster-word discovery was invoked 10 times. Then, 10 groups of cluster-word distributions $\{\vartheta^1,\vartheta^2,\ldots,\vartheta^{10}\}$ were obtained. For Algorithm 1, P_1 was set to 1000 and $P_2$was set to 1, which implied that the most representative words in each cluster tended to have the "cannot-link" relations with other words. Several examples of the must-link and the cannot-link relations uncovered from our Sina corpus are shown as follows.

- Must-link pairs: (Material, Industry), (Law, Regulation), (Challenge, Opportunity), (Catering, Service), (Apple, Cellphone), (Profit, Benefit)
- Cannot-link pairs: (Index, English), (Insurance, Art), (Warrant, Culture), (Economy, Students), (Music, Commodity), (Credit, Game)

With reference to the must-link and cannot-link examples, it can be seen that semantically related words are grouped via the must-link relations. For example, "law" and "regulation" are about legality, and "apple" and "cellphone" are related because Apple Inc. is famous with its cellphone products. In contrast, the cannot-link relations capture words that are not supposed to be semantically related. For instance, "insurance" belongs to the finance domain, while "art" is about another problem domain. Moreover, "warrant" is a kind of stock investment instrument, while "culture" is not directly related to the stock investment domain. Accordingly, these words are captured by the cannot-link relations.

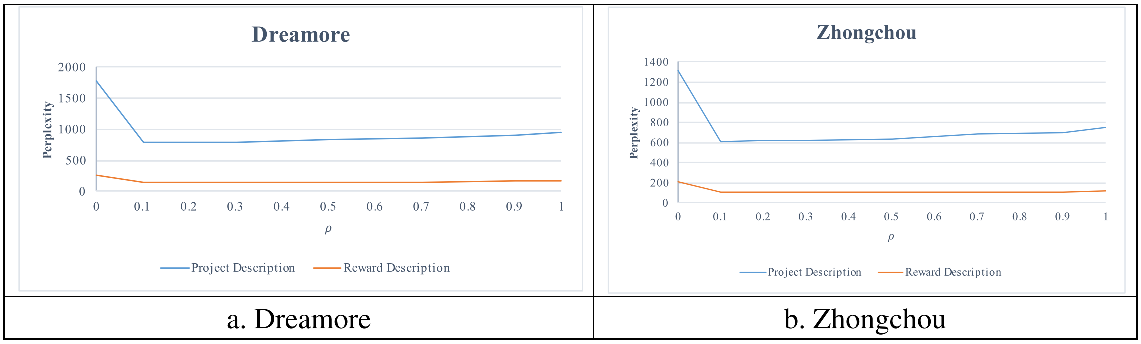
After the construction of the semantic network, the DC-LDA model was applied to the crowdfunding textual corpus. More specifically, Gibbs sampling was applied to approximate the topic-word distribution

Table 2
A confusion matrix.

		Predicted class	
		Success	Failure
Actual Class	Success	TP	FN
	Failure	FP	TN

[2] http://www.dreamore.com/.

[3] http://news.sina.com.cn/.



a. Dreamore

b. Zhongchou

Fig. 3. Parameter estimation based on the Dreamore and the Zhongchou datasets.

$\varphi$and the document-topic distribution θ of the corpus. For each textual corpus (e.g., Dreamore or Zhongchou), there are two document subsets, namely project descriptions and reward descriptions. According to the *Perplexity* measure, we empirically identified the reasonable number of topics to characterize each textual corpus. In particular, we tried different number of topics in the range between 5 and 100 for each textual corpus and observed the corresponding perplexity scores. According to our empirical tests, the reasonable number of topics should be 80 for the project descriptions and 60 for the reward descriptions for both textual corpora. Apart from establishing the topic numbers, we needed to establish the parameter $\rho\in[0,1]$. Similarly, we applied the *Perplexity* measure to identify a reasonable ρ value that optimized the predictive power of a model. Fig. 3 shows the perplexity scores with respect to different ρ values. According to our empirical test, the perplexity score is minimized when ρ falls in the range of [0.1, 0.3]. Eventually, we set $\rho=0.2$ for estimating the distribution ϕ_t based on the semantic network.

5.4. Topical feature selection

The proposed DC-LDA model can extract topical features from the project descriptions and reward descriptions to enhance the prediction of crowdfunding success. For project descriptions, the DC-LDA model can mine the topical features (i.e., topic-word distribution) that represent the inherent characteristics of the crowdfunding projects. On the other hand, the DC-LDA model can uncover the topical features representing different kinds of rewards for backing the crowdfunding projects. However, the initial topical features uncovered could be noisy, so feature selection method is applied to eliminate some redundant and irrelevant topical features. First, the attribute evaluator CfsSubsetEval [32] was employed. Then, a genetic algorithm (GA)-based feature search method GeneticSearch [47] was applied to search for a near optimal subset of topical features. Some representative topical features extracted from the project descriptions of the two crowdfunding websites are presented in Table 3.

For the Dreamore corpus, Topic 1 is about volunteer education in the remote mountainous regions in China. Since there is a trend for the Chinese to pay much more attention to improve child education in the poverty-stricken areas in recent years, backers are more likely to support crowdfunding projects about volunteer education in remote mountainous regions in China. Similarly, Topic 1 mined from the Zhongchou corpus describes technology-enabled innovative projects. Given the Chinese government's strong support for technological innovation in recent years, backers have much confidence to fund this kind of project. In addition, Topic 3 mined from Dreamore and Topic 4 mined from Zhongchou are about environmental protection and food safety, respectively. As these issues are of great concerns to people in China and the rest of the world as well, these topical features are very useful to pinpoint fund raising success as well. In contrast, the topical features mined from the reward descriptions seem not so discriminative in terms of identifying successful or unsuccessful projects. Some examples of topical features mined from the reward descriptions are shown in Table 4. For example, Topic 1 extracted from the Zhongchou corpus describes general aspects such as project, time, expense, and so on. It does not prompt for a direct relationship with successful or unsuccessful projects.

5.5. Experiments for predictive models

We evaluated the prediction performance of the proposed RF classifier and other well-known machine learning classifiers such as SVM, back-propagation neutral network (BPNN), and extreme learning machine (ELM) [48,49]. The adopted feature set included basic numerical features depicted in Table 1 and the topical features extracted from project descriptions. Table 5 reports our experimental results. Based on *Accuracy*, *Precision* and F_1, the proposed RF classifier outperforms all the baseline classifiers. However, SVM performs better than RF in terms of *Recall*. The possible reason is that the SVM classifier, which operates based on the principle of finding the optimal hyperplane to separate different classes, tends to bravely classify instances into the positive or the

Table 3
Representative topical features (project descriptions).

Topic	Dreamore	Zhongchou
1	Children, School, Students, Education, Volunteer, Volunteering, Teachers, Supporting Education, Love, Mountainous Regions, Knowledge	Product, Cellphone, Users, Intelligence, Design, Function, Technology, Brand, Science, Price, Enterprise, Price, Market, Innovation
2	Handwork, Material, Art, Color, Accessories, Work, Tools, Ancillary, Processing, Leather, Manufactured Goods	Art, Works, China, Culture, Life, Design, Create, Handwork, Spirit, Artist, Country, Painting, Dance, History
3	Protection, Rubbish, Environment, Pre-environment, Regions, Grassland, Actions, Ecology, Classifying, Advertising	School, Project, Plan, Device, Teachers, Conditions, Fund, Students, Budget, Classrooms, Volunteering, Mountainous Regions
4	Taste, Delicacy, Food, Market, Flavor, Nutrition, Consumers, Featured, Cakes, Hot-dog, Snacks, Tongue, Recipe	Standard, Food, Detection, Safety, Biology, Rice, Money, Consumers, Experiments, Country, Report, Research, Gene
5	Yunnan, Transformation, Picture, Lijiang, Photos, Albums, Postcard, Hotel, Customs, Style, Feature, Inns	Fruit, Snacks, Classmates, Undergraduate, Partner, Cost Price, Farmers, Team, Take-out, Orchardist, Price, Business Opportunity

Table 4
Representative topical features (reward descriptions).

Topic	Dreamore	Zhongchou
1	Characteristic, Time, Text, List, T shirt, Plan, Present, Friends, Gift, Internship	Project, Time, Expense, Freight, Dream, List, Community, Friends, Quota, Manner
2	Name, List, Special Version, Characteristic, Manner, Official, Project, Clothes, Member, Plan	Photo, Ambassador, Video, China, Freight, Project, Author, Activities, Expense, Time
3	Documentary, Launcher, Procedure, Team, Member, Present, Supervisory, Gift, Quota, Clothes	Inns, Freight, Discount, Friends, Experience, Opportunity, Service, Relatives, Fun, Time

negative class. As a result, the *Recall* is improved due to more positive or negative cases are found. However, among these brave classifications, there are many incorrect classifications as well. Therefore, the F_1 score of SVM is inferior to that of RF. Since the RF classifier achieved the best F_1 score for both evaluation subsets, we adopted RF as the predictive model for the rest of our experiments.

5.6. Experiments for various feature sets

We also tested the effectiveness of different combinations of features. For this series of experiments, we tried four feature sets namely, E1, E2, E3 and E4 which represented purely numerical features (as shown in Table 1), numerical features and topical features mined from project descriptions, numerical features and topical features mined from reward descriptions, and using all available features. Our experimental results are depicted in Table 6. It is obvious that the feature subset E2 (i.e., numerical features and topical features mined from project descriptions) leads to the best performance. In particular, the RF classifier can achieve F_1 scores of 78.5% and 95.15% for the Dreamore and the Zhongchou corpora, respectively. By using topical features mined from project descriptions, the percentage of performance improvement over using numerical features alone is up to 8.2% and 41.1% for the Dreamore and the Zhongchou corpora, respectively. These experimental results confirm the benefits of the topical features mined from project descriptions. However, prediction performance is not improved by using topical features mined from reward descriptions when compared to using numerical features alone. The main reason is that the reward features are noisy as shown in Table 4. The topical features of rewards cannot really help predicting project success or failure at all. By using all available features (E4), the prediction performance cannot be improved either. The reason may be that the whole feature set becomes noisy by adding the low quality topical features mined from reward descriptions. Moreover, using many features may also lead to the over-fitting problem, which causes the degradation of prediction performance.

5.7. Comparative evaluation

Table 6
Prediction performance (%) of various feature sets.

Dataset	Dreamore				Zhongchou			
Feature	E1	E2	E3	E4	E1	E2	E3	E4
Accuracy	72.00	**77.00**	70.00	63.00	71.00	**95.00**	72.00	87.00
Precision	71.75	**73.68**	68.52	62.75	76.92	**92.45**	78.95	86.00
Recall	74.00	**84.00**	74.00	64.00	60.00	**98.00**	60.00	88.00
F_1	72.55	**78.50**	71.15	63.37	67.42	**95.15**	68.18	87.13

The significance of bold is the maximum value of the prediction performance in each crowdfunding platform.

To evaluate the effectiveness of the proposed DC-LDA model for topical feature mining, we also employed the classical LDA model as the baseline for topical feature mining. For both models, the RF classifier was applied to predict fund raising success. Our experimental results are depicted in Table 7. The performance of the LDA + RF model is inferior to that of the DC-LDA + RF model depicted in Table 6. For the feature set E2 (i.e., numerical features and topical features mined from project descriptions), the LDA + RF model performs poorly for the Dreamore corpus. Based on the E2 feature set, the DC-LDA model outperforms the classical LDA model by an average of 11% in terms of F_1 score. Our experimental results confirm that the DC-LDA model is effective for mining topical features from project descriptions, and it performs better than the classical LDA model. The comparative prediction performance of these two models is plotted in Fig. 4. In fact, the DC-LDA model outperforms the LDA model across feature sets and crowdfunding datasets. The main reason of the performance improvement brought by the DC-LDA model is that the model can more effectively mine semantically valid topical features from a textual corpus. The high quality topical features are then utilized by the RF classifier to more accurately predict the fund raising successes of crowdfunding projects.

Table 5
Prediction performance (%) of various classifiers.

Dataset	Dreamore				Zhongchou			
Classifier	RF	BPNN	SVM	ELM	RF	BPNN	SVM	ELM
Accuracy	**77.00**	57.00	53.00	59.00	**95.00**	94.00	80.00	85.00
Precision	**73.68**	54.93	51.72	58.18	**92.45**	90.74	71.43	87.23
Recall	84.00	78.00	**90.00**	64.00	98.00	98.00	**100.00**	82.00
F_1	**78.50**	64.46	65.69	60.95	**95.15**	94.23	83.33	84.54

The significance of bold is the maximum value of the prediction performance in each crowdfunding platform.

5.8. Discussions

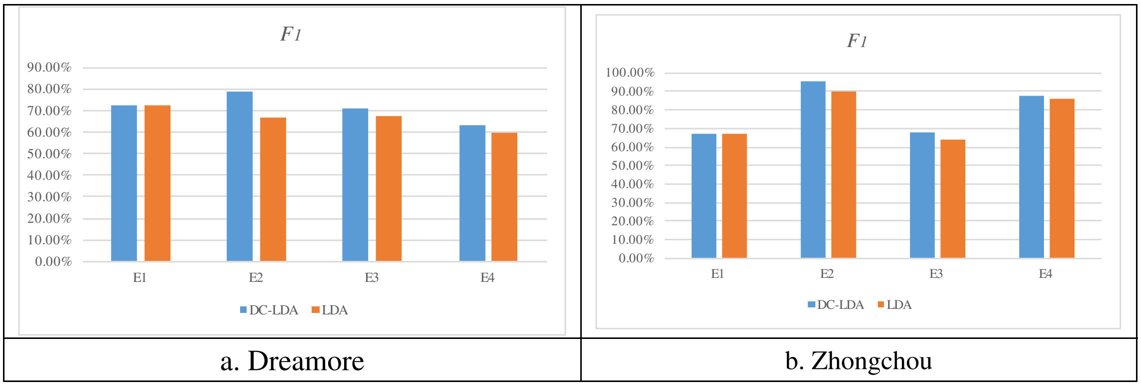
The proposed text analytics framework for crowdfunding analysis is effective. More specifically, by applying the proposed DC-LDA topic model to mine topics from the descriptions of crowdfunding projects, semantically rich topical features are obtained. These semantically rich topical features are combined with basic numerical features to build a feature set that is utilized by a RF classifier to enhance the prediction of crowdfunding success. Based on real-world data collected from two crowdfunding websites, our experimental results reveal that the proposed framework can achieve an average F_1 score of 86.8%. As a whole, our empirical results show that one important determinant of crowdfunding success is the topical features which reflect the inherent nature of crowdfunding projects. Not only do these topical features facilitate the prediction of fund raising success but they also help explaining why backers are willing to support certain crowdfunding projects (e.g., environmental protection). By mining the topical features from project descriptions, project founders can better understand the determinants of fund raising success, and hence they can better present and promote their projects via crowdfunding websites. The effective promotion of crowdfunding projects can lead to the ultimate successes of these projects.

However, the topical features extracted from reward descriptions of crowdfunding projects are relatively noisy, and these features may even hurt the prediction performance. In sum, textual features may

Table 7
Prediction performance (%) with the classical LDA model.

Dataset	Dreamore				Zhongchou			
Feature	E1	E2	E3	E4	E1	E2	E3	E4
Accuracy	**72.00**	67.00	66.00	56.00	71.00	**90.00**	69.00	86.00
Precision	**71.75**	67.35	64.81	55.00	76.92	**91.67**	75.68	87.50
Recall	**74.00**	66.00	70.00	66.00	60.00	**88.00**	56.00	84.00
F_1	**72.55**	66.67	67.31	60.00	67.42	**89.80**	64.23	85.71

The significance of bold is the maximum value of the prediction performance in each crowdfunding platform.



a. Dreamore

b. Zhongchou

Fig. 4. The comparative F_1 score: DC-LDA versus LDA.

not necessarily improve the prediction of crowdfunding success. Performance improvement is achieved only if semantically valid and relevant topical features are combined with numerical features to carry out the prediction tasks. To mine semantically valid and relevant topical features, researchers should pay attention to the textual sources (e.g., project descriptions versus reward descriptions) and the effectiveness of a topic modeling method. Moreover, our empirical tests show that different machine learning classifiers do influence the final prediction performance. By using an ensemble of weak classifiers (e.g., the RF classifier), it seems more effective than using a single powerful classifier (e.g., SVM).

6. Conclusions and future work

In the era of the social Web, crowdfunding has emerged to be one of the most important funding sources for entrepreneurs or small enterprises. While previous studies have identified some numerical features for predicting crowdfunding success, little research is conducted to explore the topical features (e.g., the latent semantics embedded in project or reward descriptions) for crowdfunding analysis. Our research work is just able to fill the aforementioned research gap. The main contributions of our work are: (1) the design of a novel text analytics framework for analyzing and predicting crowdfunding success; (2) the design of a new DC-LDA topic model to effectively mine topical features from crowdfunding project descriptions and reward descriptions to facilitate the analysis of crowdfunding projects; (3) conducting an empirical study to identify the discriminatory features that influence fund raising successes of crowdfunding projects.

Based on real-world crowdfunding datasets, our experimental results reveal that the topical features mined from project descriptions are useful to predict fund raising success. The hybrid DC-LDA and RF predictive framework achieves an average F_1 score of 86.8%. Moreover, the proposed DC-LDA topic model can effectively mine semantically valid and relevant topical features from a textual corpus, and it outperforms the classical LDA model by an average of 11% in terms of F_1 score. Surprisingly, our experimental results also show that the topical features mined from the reward descriptions of crowdfunding projects are not the main determinants of fund raising success. The managerial implications of our research work are that crowdfunding project founders can apply the proposed text analytics framework to evaluate the potentials of their proposed projects, and they should pay more attention to disseminate information about the inherent nature of their projects if they want to raise sufficient fund from the crowd to launch their projects.

Our future work will enhance the topic mining process by exploring more sophisticated topic models such as the Hierarchical Dirichlet Process (HDP) model that does not require a pre-defined number of topics. Moreover, our current study only examines five common numerical features of crowdfunding projects. Future work will incorporate more project attributes to further improve the performance of the proposed predictive model. Finally, a larger scale of experimentation will be conducted to improve both the internal validity and the external validity of our study. For instance, our current study only examines project data retrieved from two crowdfunding sites in China. We will improve the external validity of our study by analyzing project data from more crowdfunding websites in the future.

Acknowledgements

This work was supported in part by the 973 Project (No. 2012CB316205), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Nova Program (No. Z131101000413058) and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (No. 15XNLQ08). Lau's work was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Projects: CityU 11502115 and CityU 11525716), and the Shenzhen Municipal Science and Technology R&D Funding - Basic Research Program (Project No. JCYJ20140419115614350).

References

[1] A. Ordanini, L. Miceli, M. Pizzetti, A. Parasuraman, Crowdfunding: transforming customers into investors through innovative service platforms, Journal of Service Management 22 (4) (2011) 443–470.
[2] V. Etter, M. Grossglauser, P. Thiran, Launch hard or go home: predicting the success of Kickstarter campaigns, Proceedings of the 1st ACM Conference on Online Social Networks 2013, pp. 177–182.
[3] E.M. Gerber, J. Hui, Crowdfunding: motivations and deterrents for participation, ACM Transactions on Computer-Human Interaction 20 (6) (2013), 34.
[4] M. Muller, W. Geyer, T. Soule, S. Daniels, L.T. Cheng, Crowdfunding inside the enterprise: employee-initiatives for innovation and collaboration, Proceedings of the 31th ACM SIGCHI Conference on Human Factors in Computing Systems 2013, pp. 503–512.
[5] E. Mollick, The dynamics of crowdfunding: an exploratory study, Journal of Business Venturing 29 (1) (2014) 1–16.
[6] G. Burtch, A. Ghose, S. Wattal, An empirical examination of peer referrals in online crowdfunding, Proceedings of the 35th International Conference on Information Systems 2014, pp. 1–19.
[7] C.T. Lu, S. Xie, X. Kong, P.S. Yu, Inferring the impacts of social media on crowdfunding, Proceedings of the 7th ACM International Conference on Web Search and Data Mining 2014, pp. 573–582.
[8] M.D. Greenberg, B. Pardo, K. Hariharan, E. Gerber, Crowdfunding support tools: predicting success & failure, Proceedings of the 31st ACM SIGCHI Conference on Human Factors in Computing Systems 2013, pp. 1815–1820.
[9] T. Mitra, E. Gilbert, The language that gets people to give: phrases that predict success on Kickstarter, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing 2014, pp. 49–61.
[10] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.
[11] R.Y.K. Lau, Y. Xia, Y. Ye, A probabilistic generative model for mining cybercriminal networks from online social media, IEEE Computational Intelligence Magazine 9 (1) (2014) 31–43.
[12] I. Titov, R. McDonald, Modeling online reviews with multi-grain topic models, Proceedings of the 17th International Conference on World Wide Web 2008, pp. 111–120.

[13] T. Wang, Y. Cai, H.F. Leung, R.Y. Lau, Q. Li, H. Min, Product topic extraction supervised with online domain knowledge, Knowledge-Based Systems 71 (2014) 86–100.
[14] G. Xu, S.H. Yang, H. Li, Named entity mining from click-through data using weakly supervised latent Dirichlet allocation, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009, pp. 1365–1374.
[15] F. Kleemann, G.G. Voß, K. Rieder, Un (der) paid innovators: the commercial utilization of consumer work through crowdsourcing, Science, Technology & Innovation Studies 4 (1) (2008) 5–26.
[16] P. Belleflamme, T. Lambert, A. Schwienbacher, Crowdfunding: tapping the right crowd, Journal of Business Venturing 29 (5) (2013) 585–609.
[17] M. Carvajal, J.A. Garcia-Aviles, J.L. Gonzalez, Crowdfunding and non-profit media: the emergence of new models for public interest journalism, Journalism Practice 6 (5–6) (2012) 638–647.
[18] G. Burtch, A. Ghose, S. Wattal, An empirical examination of users' information hiding in a crowdfunding context, Proceedings of the 34th International Conference on Information Systems 2013, pp. 1–19.
[19] A.R. Stemler, The jobs act and crowdfunding: harnessing the power and money of the masses, Business Horizons 56 (3) (2013) 271–275.
[20] D.M. Satorius, S. Polland, Crowdfunding-what independent producers should know about the legal pitfalls, Entertainment and Sports Lawyer 28 (2) (2010) 15–17.
[21] P.Y. Kuo, E. Gerber, Design principles: crowdfunding as a creativity support tool, Proceedings of ACM Extended Abstracts on Human Factors in Computing Systems 2012, pp. 1601–1606.
[22] P. Belleflamme, T. Lambert, A. Schwienbacher, Individual crowdfunding practices, Venture Capital 15 (4) (2013) 313–333.
[23] J.S. Hui, M.D. Greenberg, E.M. Gerber, Understanding the role of community in crowdfunding work, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing 2014, pp. 62–74.
[24] G. Burtch, J. Chan, Reducing medical bankruptcy through crowdfunding: evidence from giveforward, Proceedings of the 35th International Conference on Information Systems 2014, pp. 1–19.
[25] T. Kappel, Ex ante crowdfunding and the recording industry: a model for the US, Loyola of Los Angeles Entertainment Law Review 29 (2008) 375–386.
[26] T. Aitamurto, The impact of crowdfunding on journalism: case study of Spotus, a platform for community-funded reporting, Journalism Practice 5 (4) (2011) 429–445.
[27] D. Mitra, The role of crowdfunding in entrepreneurial finance, Delhi Business Review 13 (2) (2012) 67–72.
[28] R. Wash, The value of completing crowdfunding projects, Proceedings of the 7th International Conference on Weblogs and Social Media 2013, pp. 631–639.
[29] G. Burtch, A. Ghose, S. Wattal, Cultural differences and geography as determinants of online pro-social lending, MIS Quarterly 38 (3) (2014) 773–794.
[30] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[31] P. Ravisankar, V. Ravi, G.R. Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems 50 (2) (2011) 491–500.
[32] K. Selvakuberan, M. Indradevi, R. Rajaram, Combined feature selection and classification - a novel approach for the categorization of web pages, Journal of Information and Computing Science 3 (2) (2008) 83–89.
[33] C.F. Tsai, Y.C. Hsiao, Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches, Decision Support Systems 50 (1) (2010) 258–269.
[34] D. Newman, C. Chemudugunta, P. Smyth, Statistical entity-topic models, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2006, pp. 680–686.
[35] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, Proceedings of the 17th International Conference on World Wide Web 2008, pp. 121–130.
[36] Z. Zhai, B. Liu, H. Xu, P. Jia, Constrained LDA for Grouping Product Features in Opinion Mining, Advances in Knowledge Discovery and Data Mining, Springer, Berlin Heidelberg, 2011.
[37] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study, Decision Support Systems 50 (3) (2011) 602–613.
[38] J. Kim, P. Kang, Late payment prediction models for fair allocation of customer contact lists to call center agents, Decision Support Systems 85 (2016) 84–101.
[39] T. Salles, M. Gonçalves, V. Rodrigues, L. Rocha, BROOF: exploiting out-of-bag errors, boosting and random forests for effective automated classification, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval 2015, pp. 353–362.
[40] S. Cang, H. Yu, Mutual information based input feature selection for classification problems, Decision Support Systems 54 (1) (2012) 691–698.
[41] M.A.H. Farquad, I. Bose, Preprocessing unbalanced data using support vector machine, Decision Support Systems 53 (1) (2012) 226–233.
[42] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, Proceedings of the 26th ACM International Conference on Machine Learning 2009, pp. 25–32.
[43] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2014, pp. 1116–1125.
[44] Z. Chen, B. Liu, Topic modeling using topics from many domains, lifelong learning and big data, Proceedings of the 31st International Conference on Machine Learning 2014, pp. 703–711.
[45] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, M. Welling, Fast collapsed gibbs sampling for latent Dirichlet allocation, Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2008, pp. 569–577.
[46] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.
[47] D.E. Goldberg, J.H. Holland, Genetic algorithms and machine learning, Machine Learning 3 (2) (1988) 95–99.
[48] A. Sun, E.P. Lim, Y. Liu, On strategies for imbalanced text classification using SVM: a comparative study, Decision Support Systems 48 (1) (2009) 191–201.
[49] K.P. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, 2012.

Ms. Yuan is a Ph. D. student at Department of Information Systems, City University of Hong Kong. She got her bachelor degree in Information Systems and master degree in Management Science and Engineering at School of Information, Renmin University of China. Her research interests include web mining, business intelligence and decision support systems.

Dr. Raymond Y.K. Lau is an associate professor at Department of Information Systems, City University of Hong Kong. His research interests include Big Data Stream Analytics, Text Mining, and e-Commerce. He has published over 150 research papers in international journals and conferences, such as MIS Quarterly, INFORMS Journal on Computing, Decision Support Systems, ACM Transactions on Information Systems, and IEEE Transactions on Knowledge and Data Engineering. He is a senior member of the IEEE and the ACM, respectively.

Dr. Xu is an associate professor at School of Information, Renmin University of China. He is a research fellow at Department of Information Systems, City University of Hong Kong. He got his bachelor and master degree in Mathematics at Xi'an Jiaotong University and doctor degree in Management Science at Chinese Academy of Sciences. His research interests include web mining, business intelligence and decision support systems. He has published over 70 research papers in international journals and conferences, such as Decision Support Systems, European Journal of Operational Research, Fuzzy Sets and Systems, IEEE Trans. Systems, Man and Cybernetics, and International Journal of Production Economics.