Article

Predicting the Success of Internet Social Welfare Crowdfunding Based on Text Information

Xi Chen [1,*], Hao Ding [1,*], Shaofen Fang [2] and Wei Chen [1,*]

1   School of Business and Tourism Management, Yunnan University, Kunming 650091, China
2   School of Management, Yunnan Minzu University, Kunming 650091, China; fangshaofenaha@163.com
*   Correspondence: chenxi1231@ynu.edu.cn (X.C.); dinghao_ustb@163.com (H.D.); chenwei@ynu.edu.cn (W.C.);
     Tel.: +86-0871-6503-4152 (X.C.)

Abstract: This study explored how the success of project crowdfunding can be predicted based on the texts of Internet social welfare crowdfunding projects. Through a calculation of the quantity of information and a mining of the sentimental value of the text, how the text information of the interconnected social welfare crowdfunding project affects the success of the project was studied. To this aim, a sentimental dictionary of Chinese Internet social welfare crowdfunding texts was constructed, and information entropy was used to calculate the quantity of information in the text. It was found that, compared with the information presented in the text, the fundraiser's social network factors are key in improving the success of fundraising. The sentimental value of the text positively affects the success of fundraising, while the influence of the quantity of information is represented as an inverted, U-shaped relationship. The non-ideal R-squared indices reflected that the multiple linear regression models do not perform well regarding this prediction. Furthermore, this paper validated and analyzed the prediction efficiency of four machine-learning models, including a multiple regression model, a decision tree regression model, a random forest regression model, and an AdaBoost regression model, and the AdaBoost regressor showed the best efficiency, with an accuracy R^2 of up to 97.7%. This study provides methods for the quantified processing of information contained in social welfare crowdfunding texts, as well as effective prediction on social welfare crowdfunding, and also seeks to raise the success rate of crowdfunding and thus features commercial and social value.

Keywords: Internet social welfare crowdfunding; text mining; sentiment; information entropy; machine learning models

1. Introduction

Crowdfunding refers to the services and attempts of funding a project or venture by raising small quantities of money from the public via the Internet [1]. The most common social welfare crowdfunding is aimed toward medical care [2]. Research has provided evidence that social welfare crowdfunding improves the recipient's access to medical treatment and promotes social engagement in charity [3]. Especially in developing countries with relatively incomplete medical security systems, social welfare crowdfunding is now widely chosen by people with unaffordable health expenses [4]. In undeveloped countries, social welfare crowdfunding holds the potential to fill insurance gaps and help those burdened by medical debt [5]. As a new financial service in the Internet era, social welfare crowdfunding can benefit everyone, thus becoming an important supplement to the health care insurance system.

Social welfare crowdfunding activates the social participation power of Internet platform users; however, not every project can be successful. Existing studies have examined the factors contributing to the success of social welfare crowdfunding projects from the perspectives of project environments, crowdfunding mechanisms, presentation elements,

and intrinsic motivation [6–11]. As a non-equity-based form of crowdfunding, traditional economic theories, such as transaction costs, reputation, and market design can explain factors influencing social welfare crowdfunding to some extent [12]. Some studies have analyzed the expressive characteristics of the texts of social welfare crowdfunding projects, including the picture information in the text [13], the identity clue of the subject [14], and the planning of behavior [15].These studies provide a basis and foundation to understand the fundraising factors of social welfare crowdfunding projects, but there is a lack of exploratory research that helps predict the success of projects.

In China, most social welfare crowdfunding projects acquire support through WeChat groups or Moments, which has become an infrastructure platform for social communication and information dissemination [16], and a study in the U.S. concluded that Internet crowdfunding has the potential to deepen social and health inequalities in the U.S., and the success of crowdfunding projects requires fundraisers to have both medical and media knowledge. In reality, the uneven distribution of cultural knowledge and unequal marketing ability mean that the fundraisers most in need of financial support do not necessarily have the highest probability of success [17]. Both social networking factors and the expression of fundraising texts influence the success of social welfare crowdfunding projects, but there have been no systematic examinations of factors and their importance.

Intelligent technology offers operable methods of exploiting text [18]. Text mining is a variation on a field called data mining [19]. Internet crowdfunding platform has generated a large amount of project fundraising texts, which includes fund raisers' social network elements and fundraising details. However, it is still lacking the intelligent means of text processing for social welfare crowdfunding. The intelligent processing methods of text data have seen continued expansion from single-objective solution methods to multi-objective solution methods. A series of literatures have discussed the use of multi-objective solution methods to mine the knowledge data contained in text data [20–22], including multi-objective data mining and processing of multiple documents [23]. This study analyzes how social welfare crowdfunding project information determines the success of projects based on a text mining approach, and attempts to achieve three research objectives:

RQ1. Provide methods for the quantified processing of information contained in the Chinese social welfare crowdfunding texts.

RQ2. What are the factors that can effectively predict success of social welfare crowdfunding projects? How important are these factors?

RQ3. Building prediction models for fundraising success based on machine learning approaches and validating the prediction efficiency of four machine-learning regression models, namely, a multiple regression model, a decision tree model, a random forest model, and an AdaBoost model.

This paper is organized as follows: In Section 2, the theoretical basis is analyzed, and the research method is introduced. In Section 3, the data features, structure, and text information acquiring method are introduced, as well as a description of how to construct the sentiment lexicon for social welfare crowdfunding projects and how to extract the sentimental values of the text in a targeted manner. In Section 4, based on the information extracted from crowdfunding texts, factors influencing the success of social welfare crowdfunding projects by utilizing the multiple regression model are analyzed. In Section 5, the training process is depicted, and the performance of the machine-learning prediction models is assessed. In Section 6, conclusions, applications, and several open issues are discussed, and the paper is concluded.

2. Theoretical Foundation and Research Method

2.1. Information Richness and Bounded Rationality in Decision-Making

Information richness refers to the quantity of information that can be conveyed through a communication medium in a specific period of time and the extent to which the medium enables the sender and receiver to reach a common understanding [24]. Individual emotion is reflected in the information richness of a text [25]. Information richness can efficiently

decrease uncertainty during the process of organizational decision-making [26]. In practice, there are three steps in the process of information transfer, including source-language translation, information transfer, and target-language translation. After the information has been translated from the sender, information richness may impose an impact on the information transfer as well as feedback, while the effective interaction of information may rely on the receiver's efficient acceptance and understanding of the information. Receivers of information during information processing are restricted by bounded rationality in decision-making [27].

With the fast extension of information and communication technology, diversity and richness of communication media in cyberspace have been constantly and extensively intensified. The Internet has ushered in an era of information surplus [28], and the attention of users has been highly diverted. The key factor of the effective transfer of information no longer exclusively lies in information media, and the effective processing of information by actors has become a great concern.

Information richness in cyberspace imposes an appreciable impact on users. It is the criterion to measure the quality of information disclosure on the webpage [29,30]. In addition, there is an active influence of high-quality information on a webpage on the user's trust, which will eventually affect the willingness to buy. The online introduction and evaluation with high information richness can enhance the user's perceptibility of the effectiveness, credibility, and persuasiveness of product information [31]. Based on the text structure of an Internet social welfare crowdfunding project and the information attention features of actors in the process of cyber crowdfunding, this paper divides the information contained in primitive data into two dimensions—the sentimental value and the quantity of information—and then transforms the text characteristics into numeric features, which lays a solid foundation for further probing into the connection between sentimental value, the quantity of information, and the fundraising success rate, as well as analyzing and forecasting the model efficiency of computational learning for fundraising success.

2.2. Mining of Sentiment Value in the Text

A text's sentiment refers to the intensity of positive or negative sentimental value expressed in the project's crowdfunding text. Currently, the calculation methods of sentiment for a text are based on sentiment lexicon or machine learning [32–34]. The sentiment lexicon-based sentiment calculation methods mainly rely on open-sourced sentiment lexicon or scene-based expanded sentiment lexicon [35]. A pure open-sourced sentiment lexicon applies to common scenes but not to special Internet-based crowdfunding scenes. In order to improve the accuracy of a text's sentiment, the traditional expanded sentiment lexicon-based calculation method is adopted herein for calculating the sentimental value of a text, as shown in the flow chart in Figure 1.

The jieba analyzer was adopted for text segmentation, whereby the text is segmented in an accurate way for text analysis [36]. The Stop Word Dictionary collected the prevailing stop word lists and incorporated special stop words in specific scenarios, totaling 1519 stop words. After word segmentation and stop word removal, each project text was stored in the form of strings composed of spaced words.
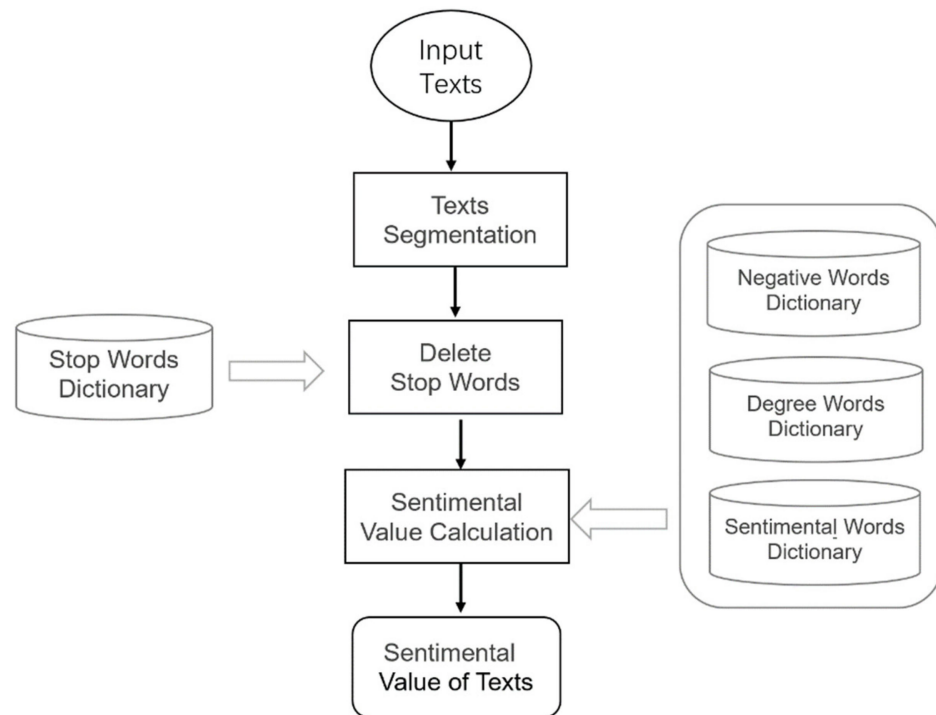
Figure 1. Calculation process of a text's sentimental value.

To calculate sentiment, it is important to construct a sentiment lexicon in the scenario surveyed, and the basic idea is to combine classic common sentiment lexicon and the Internet-based crowdfunding sentiment lexicon and eliminate repeated words for building a sentiment lexicon required in this study. To be specific, common sentiment lexicons adopted include the CNKI Sentiment Lexicon [37] and the National Taiwan University Simplified-Chinese Dictionary (NTUSD) [38]. Here, the Semantic Orientation Pointwise Mutual Information (SO-PMI) algorithm was mainly utilized to construct a scene-based sentiment lexicon. The SO-PMI algorithm consists of two parts: SO-PMI and Pointwise Mutual Information (PMI) [39]. PMI is employed for determining the probabilities of the occurrence of a certain word and the reference word, whose calculation equation is expressed below:

$$PMI = log_2 \left(\frac{P(word1, word2)}{P(word1)P(word2)} \right) \tag{1}$$

$P(word1, word2)$ refers to the joint probability, namely, the probability that *word1* and *word2* occur in the corpus simultaneously. If they are independent, $P(word1, word2) = P(word1)P(word2)$; that is to say, the total score is 1, and $PMI = 0$.

SO-PMI is used for determining the correlation between a strange word and the reference word; if it is more correlated with the positive reference word, the strange word is positive; if it is more correlated with the negative reference word, the strange word is negative. If it is equally correlated with the positive and negative reference words (but independent from these reference words), it is determined that this strange word is a neutral word, as expressed below:

$$SO - PMI(\text{word}) = \sum_{i=1}^{num(pos)} PMI(word, pos_i) - \sum_{i=1}^{num(neg)} PMI(word, neg_i) \tag{2}$$

where num(pos) refers to the total number of positive reference words, and num(neg) is the total number of negative reference words. If $SO - PMI > 0$, the strange word is considered positive; if $SO - PMI = 0$, the strange word is considered neutral; if $SO - PMI < 0$, the strange word is considered negative.

2.3. Calculation of Text Information

Text information is used for describing the information contained in a text. In this study, the information entropy of a single raising text is employed to replace the text information. The Theory of Information Entropy was proposed by Shannon in 1984 for describing the uncertainty in an information source [40]. Information entropy is often utilized as a quantitative indicator for measuring the information content of a system. The information entropy is calculated as follows:

$$H(X) = -\sum_{x \in X} p(x) log\ p(x) \tag{3}$$

Given a random variable *X*, with *n* possible outcomes x_1, x_2, ... , x_n, which occur with probability $p(x_i)$, there should be $\sum p(x_i) = 1$. Information entropy *H(X)* reflects the information content of the random variable *X* and is often represented by bits. Here, the greater the information entropy of the text is, the more information the text contains [41].

This study adopted TF-IDF (term frequency-inverse document frequency) to calculate the text information entropy in combination with the information entropy formula, of which TF is mainly used for the word frequency of internal text, while IDF is utilized for the inverse document frequency of the external text as the weight W for the information entropy of each word. IDF (inverse document frequency, opposed to document frequency) refers to feature terms (words) in a set of documents that describe a certain document feature. They can be given the corresponding weights according to their frequency in this set of documents. Special words that appear only in a few documents have larger weights than those that appear in more documents. Based on interpretation of IDF by Shannon's information theory, the higher the frequency of a feature term is in all documents, the less its information entropy will be; if a feature term appears rather intensively with higher frequency in only a few documents, it will show higher information entropy [42–44]. Therefore, IDF can be interpreted as the cross entropy of probability distribution of the keywords under a specific condition [45]. The original Shannon formula combined with word frequency in a single document only considers the information entropy of words in individual crowdfunding texts. Based on TF-IDF theory, using an IDF weighting factor, a description of information entropy of words across texts can calculate the information entropy of individual text with more details in the whole crowdfunding scenario. The following is the computational formula of information entropy for each text:

$$H(X) = -W_{x_i} \sum_{i=1}^{n} p(x_i) log\ p(x_i) \tag{4}$$

In Formula (4), *X* represents a single individual text, x_i stands for the words in the text, *n* represents the total amount of words, and $p(x_i)$ represents the word frequency of a single individual text.

$$p(x_i) = \frac{fr(x_i)}{fr(X)} \tag{5}$$

$fr(x_i)$ is the word frequency of x_i in the text, and $fr(X)$ is the value of text *X* after it is divided, and will show the total number of separated words.

$$W_{x_i} = log(\frac{D}{D_{x_i} + 1}) \tag{6}$$

W_{x_i} stands for the inverse document frequency of x_i, *D* represents the total amount of texts, and D_{x_i} stands for the number of texts containing x_i. Based on the computational formula of information entropy, the longer the single individual text, the more words contained and the larger the information entropy [46].

2.4. Predication Models for Machine Learning

2.4.1. Decision Tree Model

A decision tree, a basic classification and regression algorithm in the field of machine learning, is a tree established by choosing an appropriate strategy. It is characterized by good readability, a fast classification speed, and good understandability. The decision tree is mainly derived from the ID3 and C4.5 algorithms proposed by Quinlan [47,48], respectively, and from the CART algorithm offered by Breiman et al. in 1984 [49]. Decision trees can be classified into two categories: classification trees and regression trees. Feature selection refers to selecting a feature from numerous ones in the training data as the classification standard of the current node; different quantitative evaluation criteria can be adopted for selecting the feature, thus producing different decision tree algorithms. Among all kinds of quantitative splitting methods, the Information Theory is adopted for measuring information classification; the Information-Theory-based decision tree algorithms include ID3, CART, and C4.5, where C4.5 and CART are derived from ID3. A regression tree generally refers to the CART tree, which is also adopted herein. In most cases, compared to algebraic prediction criteria built using conventional statistical methods, a prediction tree constructed based on the CART model yields a higher accuracy, and the more complicated the data and the more variables it involves, the higher the superiority of the algorithm.

2.4.2. Random Forest Regressor

A random forest is composed of multiple decision trees that are not correlated with each other; the ultimate output of the model is jointly decided by all decision trees in the forest [50]. A certain number of samples are randomly chosen from the training set as the samples of the root node for each decision tree; while setting up each decision tree, some candidate attributes are randomly selected, and the most appropriate attribute is chosen as the splitting node. As for regression, the mean value outputted by each decision tree is taken as the result. A random forest is an Ensemble Learning method of a Bagging type; by combining multiple weak learners, the result is obtained by voting or taking the mean value so that the overall result of the model has high accuracy and generalization performance [51]. Goods results are attributed to "the randomness" and "the forest": the former endows it with an anti-overfitting capability, while the latter increases the accuracy. Weak analyzers in the random forest adopt CART trees, which are also called regression trees.

2.4.3. AdaBoost Regressor

A Boosting algorithm, a reinforcement learning method proposed by Schapire and Freund, is an important integrated learning technology that can be used for changing a weak learner, whose prediction accuracy is only slightly higher than random guessing, into a strong learner with a high prediction accuracy [52]. The AdaBoost algorithm was obtained by improving the Boosting algorithm by Freund and Schapire in 1995, which chooses weak learners with the lowest weight coefficient from the trained weak learners and integrates them into a final strong learner by adjusting the weight of the sample and the weak learner. The weak learner is trained based on the training set; the next weak learner is trained in accordance with different weight sets of the samples. The difficulty of classifying each sample decides its weight, and the classification difficulty is estimated from the outputs of the learners in the previous steps. The AdaBoost regressor is an iterative algorithm achieved by changing the data distribution. It determines the weight of each sample by judging whether the classification of each sample in every training set is correct in accordance with the overall accuracy of the previous classification. AdaBoost provides a framework in which various methods can be adopted for constructing sub-classifiers, and it allows the use of simple, weak learners without screening the features.

2.4.4. K-Fold Cross Validation and Grid-Search Gain Scheduling

Cross validation means that raw data are (generally averagely) divided into K-many groups; each subset is used as the validation set once with the other (K-1)-many subsets as a training set, thus yielding K models. The mean classification accuracy of the final validation set of the K model is adopted as a performance indicator for the classifier under K-CV [53]. Through gain-search gain scheduling, the best-performing parameter is chosen from all candidate parameters as the result. In this study, both K-fold cross validation and grid-search gain scheduling were utilized for optimizing the hyperparameters of decision tree regression, random forest regression, and AdaBoost regression, and 5-fold standards were set.

2.5. Assessment of Prediction Model Performance

In this study, the explained variance score, the mean square error, the mean absolute error, and R-square were utilized as indicators for assessing the prediction performance of the predication model [54]:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7}$$

$$MAPE(y, \hat{y}) = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{8}$$

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \overline{y} = \frac{1}{n} \sum_{1}^{n} y_i \tag{10}$$

$$R^2\_adjusted = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{11}$$

where $\hat{y}$ is the sequence of the predicted value, $\hat{y}_i$ is the predicted value of the i th sample, y_i is the expectation, n is the number of all samples, and $\overline{y}$ is the mean value of the sequence of expectations. n is the number of samples and p is the number of features.

3. Data and Features

Qschou (https://www.qschou.com (accessed on 28 December 2021) was chosen as the basic scenario and data source in this study, which is also named "Fun in Funding". Qschou is now estimated to be one of the biggest social welfare fundraising platforms; there are more than 190 million registered users, with nearly 20 billion RMB collected for 160 million Chinese families unable to cover medical expenses [55,56].

The basic data of 1249 crowdfunding projects were collected. Repeated projects were excluded, as were projects whose actual crowdfunding amount was greater than the fundraising goal, which is the targeted fundraising amount set by the fundraiser according to the consideration of potential donors [57]. Finally, 1239 crowdfunding projects were used as raw data. Each crowdfunding project consists of two kinds of data: structured and unstructured data. Figure 2 presents the data structure of a specific project on Qschou.

The data collected include structured and unstructured data. Structured data consist of "Goal", "Raised", "Help", "Forwarding", "Verify", "Success", and "Date"; unstructured data contain "project title" and "project help text". All projects adopted here are marked with "Success", but the field "Verify" in some projects is missing, and other structured data are complete; unstructured data refer to the crowdfunding text, totaling 920,000 Chinese characters. Table 1 shows the structure of the raw data.
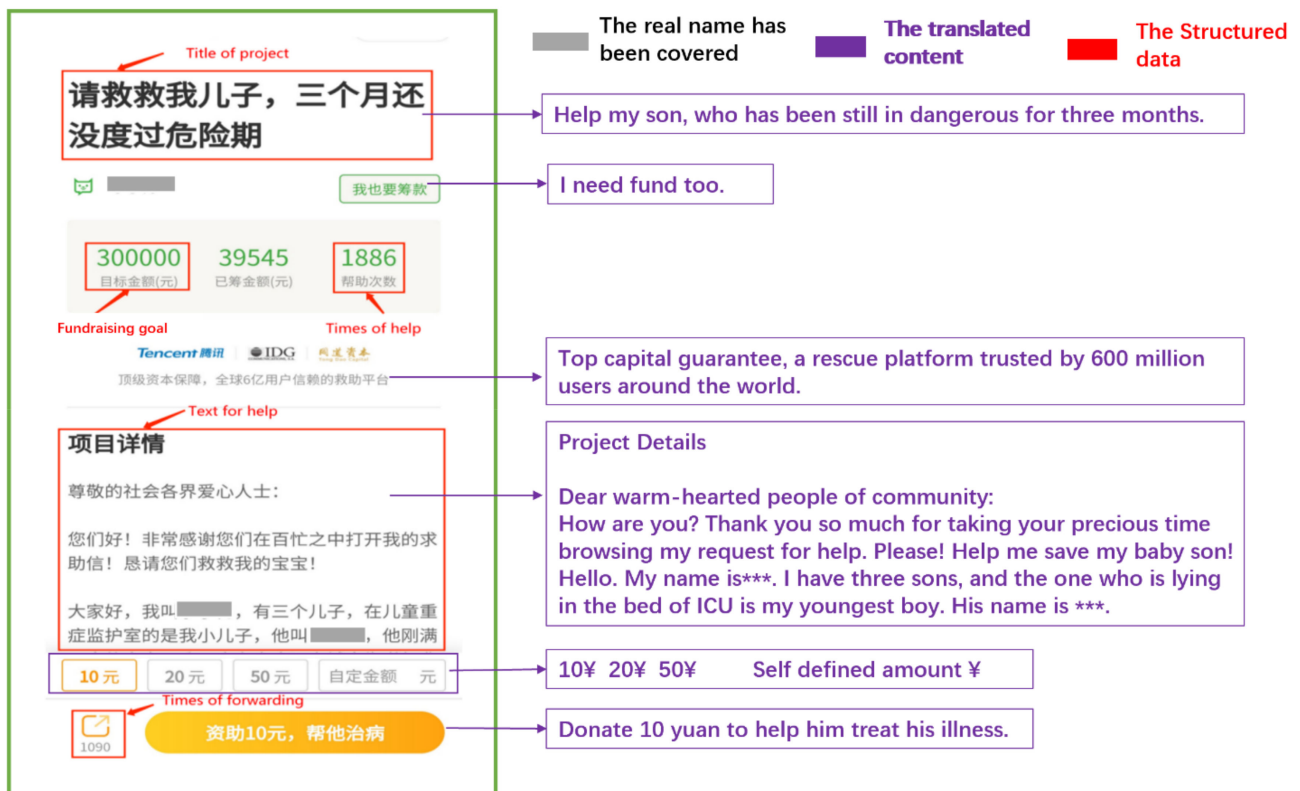
Figure 2. Sample data and basic structure of the project.

Table 1. Data structure.

No.	Goal	Raised	T_Help	T_Forwarding	N_Verify	Success	Date
1	500,000	175,727	8005	4556	193	1	2018.8
2	200,000	83,479	4394	4394	98	1	2018.9
3	500,000	119,767	2901	2901	129	1	2018.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1239	300,000	108,684	4309	1518	94	1	2019.6

No.	Title	Text
1	爸爸病危, 妈妈身亡恳请大家帮帮我吧	尊敬的叔叔阿姨您们好我叫***,现年12岁, 家住***坪上镇一个普通家庭, 我是爸爸妈妈唯一的女儿, 恳求大家帮帮我吧! 妻子在20**年*月*日车祸中身亡...
2	【急性白血病】我们一定会战胜你	尊敬的社会爱心人士你好,万不得已发起本次筹款,希望得到大家的理解与支持!我从来没想过自己会发轻松筹,事情发生的太突然了,让人猝不及防,打扰大家了,说声抱歉...
⋮	⋮	⋮
1239	孩子移植两年后,再生障碍性贫血又复发!	患有再生障碍性贫血的儿子需要骨髓移植,全家唯有8岁的姐姐和弟弟配型成功,姐姐兴高采烈盼地盼望做救弟弟的"英雄",爸妈却犹豫了,担心年幼的女儿献骨髓是否有影响孩子的爷爷站起来说:儿子女儿都是心头肉,少了哪一个都不行,如果真得需要两个孩子匹配,就放心去做,因为他们本是同根生,身体里流的都是彼此的血液...

Note: Chinese words appear in the table to present the structure of the original data, and the translations are presented in Appendix A Table A1. For the purpose of privacy protection, identity-sensitive information is blurred.

3.1. The Calculation of Information Entropy

This paper explores the overall calculation of text information entropy from the following two perspectives: (1) As for the information entropy of words in a single individual text, there are two categories of words after text segmentation. One category includes pronouns and auxiliary words with a high word frequency, such as "的" (meaning "of") and "我" (meaning "I"), but without an actual consequence; the other contains nouns and adjectives with a low word frequency, such as "肿瘤" (meaning "tumor") and "医院" (meaning "hospital"). (2) In terms of the overall text generated by all texts, the inverse document frequency of a certain word is used to indicate its significance, which is used as the weight for the computational formula of information entropy, so as to lower the interference of pronouns and auxiliary words with text information entropy. Table 2 shows the computational results of some text information entropy.

Table 2. Calculation of the information entropy of texts.

NO.	Text	Inf_Entroy	Length
1	尊敬的叔叔阿姨您们好我叫***, 现年12岁, 家住***坪上镇一个普通家庭, 我是爸爸妈妈唯一的女儿, 恳求大家帮帮我吧! 妻子在20**年*月*日车祸中身亡 . . .	10.651	566
2	尊敬的社会爱心人士你们好,万不得已发起本次筹款,希望得到大家的理解与支持!我从来没想过自己会发轻松筹,事情发生的太突然了,让人猝不及防,打扰大家了,说声抱歉 . . .	11.349	585
3	我叫***,今年**岁,**省***村人八月底因有半个月身体难受没劲,去医院检查,检查后确诊是尿毒症,这个消息就像晴天霹雳,我们不敢相信,辗转去了很多家医院反复检查,最后还是确诊为尿毒症 . . .	11.704	476

Note: Chinese words appear in the table to present the structure and features of the original data, and the translations are presented in Appendix A Table A2. For the purpose of privacy protection, identity sensitive information is blurred.

3.2. Building the Sentimental Dictionary

In the application of machine leaning technologies, text features and numerical features are different [58,59]. In this study, the information from the raw data includes two dimensions: sentiment and information content, so the text features were converted into two dimensions of numerical features; this study further explored the relationships between the sentiment and information content of the text and the success rate of the crowdfunding project and lay a foundation for understanding the efficiency of the machine-learning models that analyzed and predicted the crowdfunding success rate.

In the early stage, three postgraduates specializing in Internet research sorted out the segmented words, and three other postgraduates focusing on Internet governance chose 50 positive sentiment words and 50 negative sentiment words in the scenario herein based on five topics, including "disease", "family", "mood", "hope", and "money". Under each topic, there were 10 positive sentiment words and 10 negative sentiment words, as shown in Table 3. These 100 seed words were used as reference words in the SO-PMI algorithm, and more than 5000 positive and negative sentiment words were ultimately expanded in this scenario. For the purpose of the accurate segmentation of sentiment words and the scientific calculation of sentiment, the sentiment lexicon required herein was constructed by combining positive and negative sentiment words with an SO-PMI score ranking the top 1000 and the integrated common sentiment lexicon and by eliminating repeated words. The sentiment lexicon constructed based on inter-information consists of 1000 positive words and 1000 negative words. After integrating this with the common sentiment lexicon and

eliminating repeated words, 9047 positive sentiment words and 13,251 negative sentiment words were included. Thus, the sentiment lexicon proposed was constructed.

Table 3. Sentimental seed words.

Emotion	Seed Words
Positive	康复,好转,健康,良性,战胜,一线生机,治好病,奇迹,维持,渡过难关,阖家幸福,安享晚年,养育之恩,美满,和睦,莫大,感谢,良好,坚强,勇敢,相信,变好,笑容,自信,温暖,帮助,跪谢,必定,恳请,爱心,好心人,活下去,努力,全力,绝不放弃,点亮,伸出,顺利,转发,筹集,捐助,聚沙成塔,积少成多,报销,富裕,众筹,坚持,加油,好起来,可以
Negative	严重,恶性肿瘤,手术,化疗,白血病,大病,复发,急性,抢救,疼痛,病魔,破碎,完了,无奈,晴天霹雳,不幸,着急,伤心,只能,放弃,没办法,无助,倾尽全力,费用,积蓄,高昂,花光,天文数字,用光,返贫,难于登天,压垮,单亲,年迈,艰辛,遗弃,自责,失去,走投无路,噩耗,彷徨,抛弃,缺乏,音讯全无,可怜,束手无策,薄弱,家徒四壁,微薄,负债

Note: Chinese words appear in the table to present the features of the sentimental seed words and the translations are presented in Appendix A Table A3.

By calculating the traditional sentiment of the text and assuming that the sentiment satisfies a linear superposition, we endowed each positive sentiment word with a weight of 1 and each negative sentiment word with a weight of −1. Afterwards, word segmentation was implemented on the sentence: if the word vectors included corresponding words after the segmentation, the forwarding weight was added. Negative words and adverbs of degree had special discrimination rules: negative words led to the opposite weight, which means the value weight should be multiplied by –1; meanwhile, the most extreme adverb of degree doubled the weight, and the weight of adverbs of moderate degree is 1.5. The sentiment of the text was determined based on whether the total weight was positive or negative. The degree and negative word lists proposed herein refer to three levels and their weight settings according to the classification degree of Chinese adverbs [60]. The calculated sentiment is shown in Table 4.

Table 4. Negative word and degree adverb dictionary.

Type	Words	Weight
Negative Word Dictionary	不曾,未必,没有,不要,难以,不是,没,未,别,莫,勿,不必,白,非,无需,并非,毫无,绝不,休想,永不,未尝,从不,从未,尚未,从没,绝非,切莫,绝不,毫不,禁止,忌,拒绝,杜绝,没有	−1
Degree Adverb Dictionary	百分之百,倍加,不堪,不得了 ...	2
	不过,不胜,惨,出奇,大为,实在 ...	1.5
	大不了,更,还要,远远 ...	1

Note: Chinese words appear in the table to present the features of negative word and degree adverb dictionary, and the translations are presented in Appendix A Table A4.

Based on the three levels of reference adverb classification and weight setting [61], the sentimental value features were defined and calculated, as shown in Algorithm 1.

Algorithm 1: Calculating Sentimental Value of Features

Inputs: Word segmentation, string after removing stop word (with space as interval), negative words dictionary, degree words dictionary, sentiment dictionary
Algorithm:
Wordlist←string after removing stop word
no_dict_list← read key:word,Value:weight←
neg_dict_list←read key:word,Value:weight
sen_dict_list←read key:word,Value:weight
score = 0
weight = 1
for word in wordlist
If (word in sen_dict_list.key and not in deg_dict_list.Key and not in no_dict_list) do
    sen_dict.key←word.index
    sen_dict.value ← word.weight
Else if(word in No_dict_list.key) do
    no_dict.key←word.index
    no_dict.value ← word.weight
Else if(word in deg_dict_list.Key) do
    deg_dict.key←word.index
    deg_dict.value ← word.weight
    end
for i running from 1 to len(wordlist):
    If (i in deg_dict.key) do:
    weight←weight*deg_dict.value
    else If (i in no_dict.key) do:
    weight←weight*no_dict.value
    else If (i in sen_dict.key) do:
    score←score + w*sen_dict.key
    End
Output:score

4. Predictor Analysis

Multiple linear regression models are often used in prediction, which is represented by the relationship between the dependent variables and a set of predictor variables [62,63]. Based on multiple linear regression, this paper carries out predictions for the success of Internet social welfare crowdfunding projects. Taking the fundraising ratio of charitable crowdfunding projects as the dependent variable Y, and the sentimental value as well as the quantity of information from the numeric data and text mining as independent variables, the influence of the independent variables on the fundraising ratio of the dependent variable was analyzed by setting up multiple linear regression equations. The regression models in this paper are as follows:

$$\text{M1: } Y = a_{10} + a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}C_{01} + a_{17}C_{02} + \varepsilon \quad (12)$$

$$\text{M2: } Y = a_{20} + a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4 + a_{25}X_5 + a_{26}X_6 + a_{27}C_{01} + a_{28}C_{02} + \varepsilon \quad (13)$$

$$\text{M3: } Y = a_{30} + a_{31}X_4 + a_{32}X_5 + a_{33}C_{01} + a_{34}C_{02} + \varepsilon \quad (14)$$

$$\text{M4: } Y = a_{40} + a_{41}X_4 + a_{42}X_5 + a_{43}X_6 + a_{44}C_{01} + a_{45}C_{02} + \varepsilon \quad (15)$$

Thereinto, Y represents the fundraising success ratio, X_1 stands for the forwarding number, X_2 is the verifying number, X_3 is the fundraising goal, X_4 represents the text's sentimental value, X_5 is the quantity of information, X_6 is the quadratic term of the information quantity, C_{01} is the sentimental value of the title, C_{02} represents the length of the title, a are the coefficients, and ε is the error term. X_5 and X_6 go are calculated based on the formulas and processes defined in "Section 2.3".

This paper builds four multiple regression models for different purposes. Model 1 and Model 2 are full models, the independent variables of which include the forwarding

number (X_1), the verifying number (X_2), the fundraising goal (X_3), the text's sentimental value (X_4), the quantity of information (X_5), the sentimental value of the title (C_{01}), and the length of the title (C_{02}). Compared with that of Model 1, there is an added quadratic term of the information quantity (X_6) in Model 2. These can fulfill two purposes. On the one hand, it can analyze the influence of all independent variables on the dependent variable; on the other hand, it can explore the connection between the quantity of information and the fundraising ratio.

Without regard to fundraising goal and other variables, Model 3 and Model 4 center on the influence of the text and the title on the fundraising ratio. The independent variables of Model 3 only include the sentimental value (X_4), the quantity of information (X_5), the sentimental value of title (C_{01}), and the length of the title (C_{02}), while there is an added quadratic term for the information quantity (X_6) in Model 4. In general, comparing the influence analysis of global variables on dependent variables in Model 1 and Model 2 with Model 3 and Model 4 adds prominence to the impact of the forwarding number (X_1), the verifying number (X_2), and the fundraising goal (X_3) on the fundraising ratio; in the comparison between Model 3 and Model 4, taking the text's sentimental value (X_4), the information quantity (X_5), the sentimental value of the title (C_{01}), and the length of the title (C_{02}) as control variables, and the quadratic term of the information quantity (X_6) as the independent variable, this paper further investigates the impact of the quantity of information and the fundraising ratio. The descriptive statistics of variables are shown in Table 5.

Table 5. Descriptive statistics.

Variables	(1) N	(2) Mean	(3) sd	(4) Q1	(5) Q2	(6) Q3	(7) Min	(8) Max
X_1	1239	1454	2787	469	1057	1925	0	53,086
X_2	1209	79.03	224.6	37	60	92	0	4494
X_3	1239	366,826	435,061	180,000	300,000	490,000	20,000	500,000
X_4	1239	−30.95	23.77	−47	−30	−14	−126	30
X_5	1239	11.71	0.669	11.34	11.76	12.15	8.192	13.55
X_6	1239	137.60	15.29	128.40	138.17	147.53	67.12	183.50
C_{01}	1239	−0.26	2.75	2	0	−2	−14	9
C_{02}	1239	21.81	8.82	17	20	25	4	79
Y	1239	0.29	0.21	0.11	0.27	0.43	0.00	1

The frequency of the verifying number (X_2) was 1209, and that of other variables was 1239, with no omission. The minimum value of the forwarding number (X1) and the verifying number (X2) was 0, and the maximum values were 53,086 and 4494, respectively. Since there were relatively small mean values and a relatively large standard deviation for the two variables, there was a great discrepancy and strong heterogeneity in the sample. With the minimum value of the fundraising goal (X3) reaching 20,000, a maximum value of 500,000, and a standard deviation of 435,061, there was a great discrepancy among the expectations of various fundraising projects, and a majority of project originators expected too much. The mean value of the text's sentimental value (X4) was −30.95, the standard deviation of which was 23.77, indicating that there was negative text sentiment. There was a relatively small standard deviation for the information quantity (X5), which indicated that the difference in the information quantity was also relatively small. With a minimum fundraising ratio (Y) of 0, a maximum value of 1, a mean value of 0.29, and a standard deviation of 0.21, the fundraising ratio was generally low, and there was a minority of fundraising projects with an extremely low or extremely high fundraising ratio. Table 6 describes the Pearson correlation analysis on all input and output variables.

Table 6. Pearson correlations of variables.

Variables	X_1	X_2	X_3	X_4	X_5	X_6	C_{01}	C_{02}
X_1	1							
X_2	0.060 **	1						
X_3	0.095 ***	0.065 **	1					
X_4	0.0330	−0.0200	0.0140	1				
X_5	−0.051 *	0.0100	−0.053 *	−0.457 ***	1			
X_6	−0.063 **	−0.00500	0.058 **	−0.0390	0.139 ***	1		
C_{01}	0.059 **	−0.00600	0.0210	0.166 ***	−0.144 ***	−0.092 ***	1	
C_{02}	−0.0440	−0.0240	−0.0410	0.0190	0.069 **	0.088 ***	−0.313 ***	1
Y	0.196 ***	0.076 ***	−0.139 ***	0.069 **	−0.048 *	−0.066 **	0.0190	−0.0210

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 7 shows the fitting results of the regression model. It can be seen, based on Model 1 and Model 2, that the goal has a significant negative influence on the raising rate, which is possibly because potential donors are more willing to help those projects whose raising goals can be easily realized and will gain more satisfaction from it [64]. The forwarding number and the verifying number exert a positive influence on the raising rate; they are often strongly and positively correlated with the social network of the fundraiser and are key factors that determine the success of crowdfunding. The sentiment of a text has a relatively significant positive influence on the success of crowdfunding; it is not so significant because, when donating money, potential donors consider the social relations of the beneficiary but not the actual situation.

Table 7. Results of the multiple regression model.

Variables	Model 1	Model 2	Model 3	Model 4
X_1	0.2041 *** (0.0277)	0.2019 *** (0.0277)	— —	— —
X_2	0.0713 *** (0.0275)	0.0711 *** (0.0275)	— —	— —
X_3	−0.1644 *** (0.0276)	−0.1615 *** (0.0277)	— —	— —
X_4	0.0586 * (0.0312)	0.0312 * (0.0312)	0.0611 * (0.0323)	0.0632 * (0.0323)
X_5	−0.0199 (0.0311)	−0.0144 (0.0313)	−0.0187 (0.0321)	−0.0100 (0.0323)
X_6	— —	−0.0385 (0.0280)	— —	−0.0607 ** (0.0288)
C_{01}	−0.0078 (0.0294)	−0.0099 (0.0295)	−0.0004 (0.0304)	−0.3339 (0.0304)
C_{02}	−0.0187 (0.0291)	−0.0164 (0.0291)	−0.0206 (0.0301)	−0.0170 (0.0301)
Constant	−0.0000 (0.0274)	−0.0000 (0.0274)	−0.0000 (0.0284)	−0.0000 (0.0283)
Adjusted R-square	0.0682	0.0688	0.0023	0.0051
Number of project	1239	1239	1239	1239

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; numbers in () are standard errors.

In Model 1 and Model 2, the text information, the sentiment of the title, and the length of the title do not significantly affect the raising rate, but the coefficient of text information squared in Model 4 is −0.0607, which is significant when $p < 0.05$, suggesting that text information has an inverted, U-shaped relationship with the raising rate; when writing the text, the fundraiser should control the length of the text to increase the raising rate, such that it should be neither too long nor too short [65].

As it is shown in Table 8, tolerance of both explanatory variables and control variables is larger than 0.1 while variance inflation factor (VIF) is smaller than 10. Independent variables show no multicollinearity.

Table 8. Multicollinearity test parameters.

Variables	Model1	Model2	Model3	Model4
	Tolerance (VIF)	Tolerance (VIF)	Tolerance (VIF)	Tolerance (VIF)
X_1	0.983 (1.020)	0.979 (1.020)	— —	— —
X_2	0.993 (1.010)	0.993 (1.010)	— —	— —
X_3	0.986 (1.010)	0.980 (1.020)	— —	— —
X_4	0.774 (1.290)	0.773 (1.290)	0.774 (1.290)	0.773 (1.290)
X_5	0.780 (1.280)	0.767 (1.300)	0.783 (1.280)	0.770 (1.300)
X_6	— —	0.963 (1.040)	— —	0.971 (1.030)
C_{01}	0.869 (1.150)	0.866 (1.150)	0.870 (1.150)	0.868 (1.150)
C_{02}	0.891 (1.150)	0.888 (1.130)	0.893 (1.120)	0.890 (1.120)
Mean	0.885 (1.130)	0.893 (1.120)	(0.892) 1.210	(0.892) 1.210

The adjusted R-squared indices are small, and at least one coefficient of the models is significantly different from 0. There are three possible reasons leading to such a phenomenon. One is that certain important variables might not be included in the model, the second is that there might be a large amount of random data interference in the original dataset [66], and the third is that the model failed to fit the actual data distribution well [67]. Such a relationship between independent and the dependent variables may also be very important, even though it may not explain a large amount of variation in the response [66]. This study aims to build methods to predict the success of Internet social welfare crowdfunding projects based on text information. Through text information extraction, we identified the factors that significantly contribute to project success predictions. The non-ideal, R-squared indices reflected that the multiple linear regression model does not perform well regarding the prediction issue discussed in this study. Therefore, we need to continue looking for prediction models with excellent performance with the help of machine learning.

5. Machine-Learning Prediction Model Performance

In order to effectively predict the raising rate, the abovementioned forwarding number, verifying number, goal, text sentiment, and text information were adopted as independent variables, and the raising rate was the dependent variable; different machine-learning regressors were adopted for prediction, including multivariate linear regression, the decision tree regressor, the random forest regressor, and the AdaBoost regressor.

The processed data sets herein were divided for training and testing prediction models by a ratio of 3:1. To be specific, the datasets for the training models were classified into training sets and validation sets based on five-fold standards; a training set is mainly used for training models, while a validation set is mainly adopted for determining network structure or controlling the complexity parameters of models; a testing set is often employed for testing the performance of trained models.

Table 9 shows how resampling techniques can be used for five-fold cross-validation on training data to obtain optimal hyper-parameters, as well as the mean value and standard deviation of the prediction performance of validation sets on trained models. For the AdaBoost regressor compared to other models, MSE = 0.040, MAE = 0.072, MAPE = 0.222, and R^2 = 0.961, and the AdaBoost regressor had a smaller prediction error and an excellent

prediction performance, followed by the decision tree regressor, the random forest regressor, and linear regression in sequence. Each model has the following optimal hyper-parameters: DecisionTree (criterion = mse, max_depth = 30, min_samples_leaf = 2), RandomForest (max_depth = 1000, min_samples_leaf = 1, max_features = None, n_estimators = 800, bootstrap = True), and AdaBoost (DecisionTreeRegressor (criterion = 'mse', max_depth = 30, min_samples_leaf = 2), n_estimators = 500, random_state = 0, learning_rate = 0.2).

Table 9. The prediction performance of validation data of balance data.

	N	**MSE** Mean (SD)	**MAE** Mean (SD)	**MAPE** Mean (SD)	R^2 Mean (SD)
Linear Regression	5	0.995 (0.058)	0.788 (0.011)	1.310 (0.210)	0.031 (0.040)
Decision Tree Regressor	5	0.145 (0.105)	0.165 (0.075)	1.941 (0.571)	0.857 (0.103)
Random Forest Regressor	5	0.115 (0.059)	0.230 (0.057)	1.750 (0.278)	0.887 (0.058)
AdaBoost Regressor	5	0.040 (0.044)	0.072 (0.058)	0.222 (0.170)	0.961 (0.043)

The random forest regressor and AdaBoost regressor both adopt a CART-based algorithm and can identify the variables that can effectively reduce impurities.

Figure 3 displays the arithmetic mean value of variable importance after five instances of repeated training. In the random forest regressor and AdaBoost regressor, the goal, forwarding number, and verifying number are all important factors that affect the raising rate, and the forwarding number is the most important one. The project initiator may make use of its social links to engage more people in forwarding crowdfunding projects through online channels and hopes that more insiders can prove the authenticity of the project. This, in turn, proves that the project initiator's social network does have a decisive role in the actual crowdfunding activities, and that the preset goal also exerts an important influence on the raising rate. The text sentiment, text information, the sentiment of the title, and the length of the title do not have a significant influence on the raising rate, suggesting that the initiator's social relations are more important than the project contents.
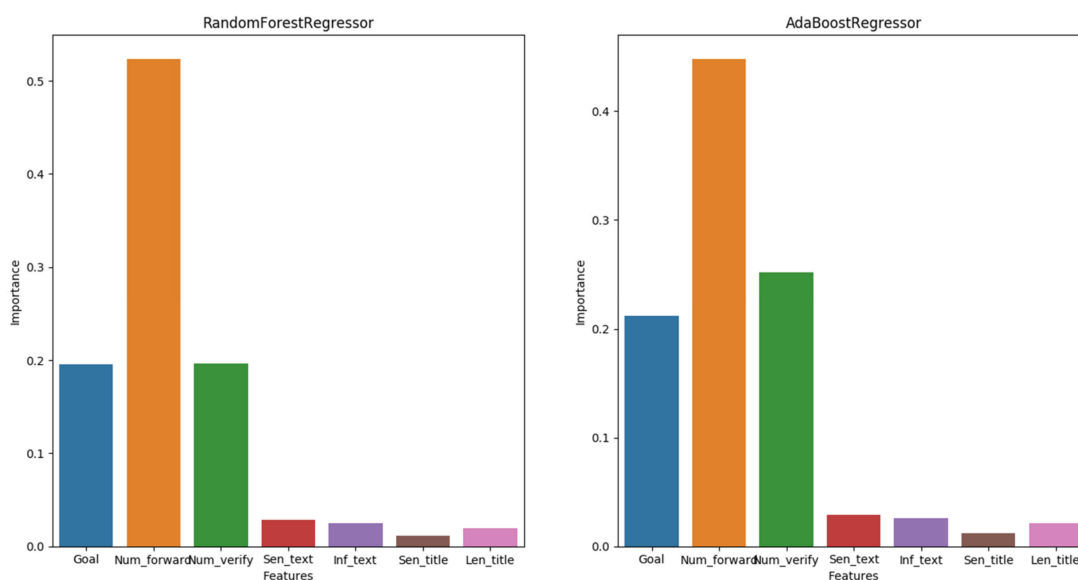

Figure 3. Arithmetic mean value of variable importance with five instances of repeated training.

Table 10 demonstrates the evaluation results of the prediction performance of the data tested by the four models using the optimal hyper-parameters. The AdaBoost regressor offers the best prediction performance; among its assessment indicators, MSE = 0.023, MAE = 0.042, MAPE = 0.090, and R^2 = 0.977. The smaller the MSE, Mae and MAPE are, the better the model performs.

Table 10. The prediction performance of test data.

	N	MSE	MAE	MAPE	R^2
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Linear Regression	5	0.987 (0.133)	0.785 (0.023)	1.373 (0.229)	0.029 (0.080)
Decision Tree Regressor	5	0.112 (0.025)	0.114 (0.014)	0.918 (0.621)	0.889 (0.026)
Random Forest Regressor	5	0.120 (0.014)	0.225 (0.008)	1.505 (0.428)	0.881 (0.012)
AdaBoost Regressor	5	0.023 (0.005)	0.042 (0.006)	0.090 (0.041)	0.977 (0.005)

According to R^2, the AdaBoost regressor shows the best prediction performance, followed by the decision tree regressor, random forest regressor, and linear regression in sequence. In particular, linear regression has the poorest prediction performance, and further showed that multiple linear regression does not serve as a good prediction model in this study. Although multiple linear regression could help us verify factor importance, the low R^2 indicates that such factors are likely to have a nonlinear and complex relationship with project success. Therefore, it is necessary to use machine learning models to realize the effective improvement of text information indices on project success prediction, proving the speculation of the above raising rate prediction model that the selected, nonlinear machine-learning model improves the poor fitting results of the multivariate regression model.

Figure 4 shows the differences between the predicted and expected values of the testing data on different models, which displays the prediction performance of four models in a more intuitive way.
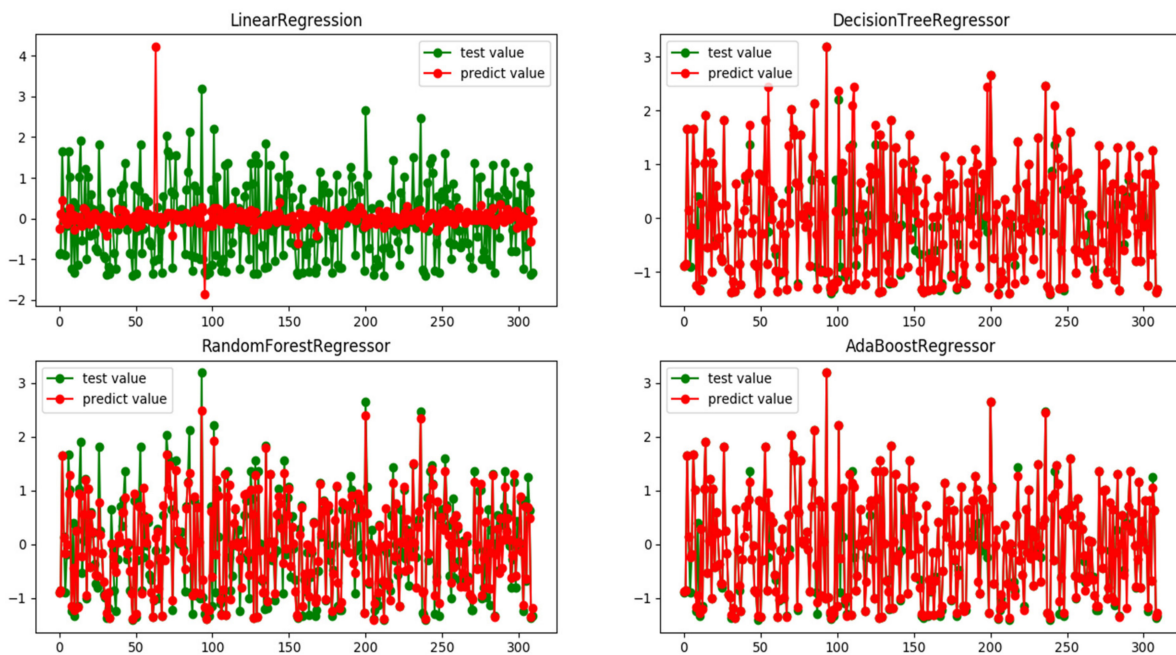

Figure 4. Predicted and expected values of testing data on different models.

6. Conclusions, Applications, and Future Work

6.1. Conclusions

Based on the method of text mining, information in the text dataset of Internet social crowdfunding projects in China, including expression factors, such as information entropy, sentimental value, and captions, as well as socializing factors, such as the number of people forwarding and the number of people confirming, was extracted to predict the success of project fundraising. For research purposes, a sentiment lexicon of text used in Internet social crowdfunding was constructed; moreover, information entropy was introduced to calculate the information quantity of the text. According to the research, compared with the information presentation of the text itself, the social networking elements of the fundraiser are key to the success of fundraising. The sentimental value of the text has a positive effect on the success of fundraising, while the effect of information quantity thereon is in a reverse U-shape. The positive effect of sentimental value supports the concerns of Berliner and Kenworthy (2017); a strong sentimental description in a fundraising text will increase the likelihood of receiving donations. This gives fundraisers with sentimental description skills an edge.

Here, information entropy was introduced to calculate the function of information quantity in predicting the success of fundraising for crowdfunding projects for the public good. It was noticed that the effect of information quantity on fundraising success is in a reverse U shape. This verifies the applicability of the findings of Simon (1972) about the limited information handling capacity of humans to the issue of fundraising for Internet social crowdfunding projects. Although the abundance of information presented to donors can help them better understand those in need of help, because of the limitations of human information processing, greater information quantity will not necessarily lead to a higher success rate; an appropriate supply of information is more beneficial to the success of such projects.

According to this research, compared with the information presentation in the text itself, socializing factors, such as the number of people confirming and the number of people forwarding, have more significant effects on the success of fundraising for Internet crowdfunding projects for the social good. Social networks are the main platforms for disseminating social welfare crowdfunding information, and social factors play an important role in the success of such projects [68,69]. This is related to the following two characteristics of such projects: being non-profit and being low in quantity. More donors chose to donate money out of love and the social validation of the person forwarding such fundraising information. In this era of information overload and distracted attention, donors will not spend too much time or energy carefully reading the text used for non-profit crowdfunding. In such a context, socializing factors seem particularly important.

Furthermore, the predictive efficiency of four machine learning models, namely, the multiple regression model, the decision tree regression model, the random forest regression model, and the AdaBoost regression model, was verified and analyzed. It was noticed that, aside from linear regression, the other three models showed good predictive efficiency. Among them, the AdaBoost regressor showed the best efficiency, with an accuracy of up to 97.7%.

6.2. Applications

The extending research and application value of intelligent method adopted by this study provides methods and tools for processing the text message of social welfare crowdfunding.

The crowdfunding text message processing adopted by the research is divided into two parts. One is the calculation of single crowdfunding text's emotional value. To acquire specific, accurate, and significant emotional value, we adopt an emotional value calculation method based on an emotional dictionary, with the critical point being the scientific selection of emotional seed words and defining the weight of words of degree appropriately. The other is the calculation of the information quantity in the text. With the view of quantifying the information content contained by single crowdfunding text and

further analyzing the influences of the information quantity contained in the descriptive text of crowdfunding on the success rate of crowdfunding, we combined the information entropy formula with the TF-IDF method under the guidance of information entropy theory and in accordance with TF-IDF theory. The information quantity contained by crowdfunding text is thus calculated.

The text information processing method in this study features value in three aspects: To begin with, this study provides methods for the quantified processing of information contained in the social welfare crowdfunding texts. Specific and operable tools are presented for the calculation of the information quantity and emotional value contained in the text. For instance, the special emotional dictionary for social welfare crowdfunding text has been established. These methods would benefit the study on the descriptive text of social welfare crowdfunding project and improve the shortages of current study and application in terms of a simple text processing method and serious missing text information. Then, the method of this study could be expanded to other scenarios for text information processing. In particular, if the qualification operation for text information is needed for the long descriptive text of a special sector, the method adopted by this research would have reference significance. Furthermore, the method adopted in this study is simple and effective. Current popular methods, such as the method adopting a neural network, have a high processing cost. It solves problems in new research scenarios based on classical text information processing theory, which provides extensible thinking for the future study.

In addition, this study provides a method for the effective prediction of social welfare crowdfunding projects and seeks to raise the success rate of crowdfunding, and thus features significant commercial and social value.

6.3. Future Work

In this paper, based on the text information in project datasets collected on social welfare Internet crowdfunding projects in China, the success rate of such projects is predicted. Future research may extend this method to datasets from Internet crowdfunding platforms in other countries. In addition, there are many other factors to be included in the prediction model: external factors, such as the economic environment and platform policy, and internal factors, such as the personal characteristics of the help-seekers and the characteristics of potential donors. Research on these aspects can help platforms better grasp success factors, better distinguish projects more likely to succeed, and thus better manage their projects.

Appendix A

Translations for the Chinese Text information in Tables 1–4.

Table A1. Data structure.

No.	Goal	Raised	T_Help	T_Forwarding	N_Vefify	Success	Date
1	500,000	175,727	8005	4556	193	1	2018.8
2	200,000	83,479	4394	4394	98	1	2018.9
3	500,000	119,767	2901	2901	129	1	2018.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1239	300,000	108,684	4309	1518	94	1	2019.6

No	Title	Text
1	Dad is critically ill and Mom was dead. Please give me a help hand.	Dear uncles and aunts. My name is ***, 12 years old, living in a ordinary family in Pingshang Town, I am the only daughter of the family and I implore everyone to help me! Wife died in a car accident on 20** . . . .
2	[Acute leukemia] That will be defeated definitely	Dear social benevolent personage, I have no choice but to initiate this fundraising, hope to get everyone's understanding and support! Never thought that I would make a QingSongChou 'cause it happened too suddenly, which caught me off guard and bothered everyone. Sorry about this situation . . .
⋮	⋮	⋮
1239	Two years after this child's transplant, the aplastic anemia recurred!	The son with aplastic anemia needs a bone marrow transplant, and only the 8-year-old sister and brother in the family have successfully matched. The sister is eagerly looking forward to being the "hero" who saves her brother as the parents hesitated, worrying about whether the young daughter's bone marrow donation would impact her healthy. At the moment, grandfather stood up and said: boy and girl are both significant that anyone shouldn't be ignored. Go for it as they are born from one bloodline which is the best convenience they are the slblings . . .

Table A2. Calculation of the information entropy of texts.

NO.	Text	Inf_Entroy	Length
1	Dear uncles and aunts. My name is ***, 12 years old, living in a ordinary family in Pingshang Town, I am the only daughter of the family and I implore everyone to help me! Wife died in a car accident on 20**	10.651	566
2	Dear social benevolent personage, I have no choice but to initiate this fundraising, hope to get everyone's understanding and support! Never thought that I would make a QingSongChou 'cause it happened too suddenly, which caught me off guard and bothered everyone. Sorry about this situation . . .	11.349	585
3	I am ***, ** years old, coming from ** province *** village. I went to the hospital and had a checkup when I felt uncomfortable at the late August and was diagnosed as uremia later. The news was like a bolt from the blue. My family couldn't believe it and then I had repeated checks in other hospitals that finally diagnosed as uremia . . .	11.704	476

Table A3. Sentimental seed words.

Emotion	Seed Words
Positive	Rehabilitation, improving, healthy, benign, overcome, a chance to cure disease, treat the disease, miracle, maintain, survive, a happy family, enjoy one's old age, gratefulness for nurturing, perfect, peace, great, thanks, favorable, strong, brave, do believe, turn good, smile, self-confidence, warm, help, grateful, surely, beg, love, virtuous people, survive, do my best, do the best, never give up, light up, giving a help hand, successfully, spread it around, raise, donate, many a little makes a mickle, many a little makes a mickle, reimburse, wealthy, crowdfunding, persevere in, come on, get better, could
Negative	serious, malignant tumor, surgery, chemotherapy, leukaemia, a serious illness, relapse, acute, rescue, pain, disease, broken, going bad, helpless, the news come like a bolt from the blue, unfortunately, being burnt with anxiety, sadness, only, give up, have no means, helpless, try my best efforts, costs, savings, high, run out, vast sums, wipe out, live in poverty again, be as difficult as climbing up to heaven, crushed, single parent family, elderly, hardship, abandoned, self-blame, loss, helplessness, bad news, loss, abandonment, lack off, out of touch, poor thing, have no ideas, weakness, destitution, sincere, debt

Table A4. Negative word and degree adverb dictionary.

Type	Words	Weight
Negative Word Dictionary	never, for not sure, have not, don't, can't, it's not, do not, did not, do not, do not, do not, need not, in vain, have to, need not, it's not that, without, not at all, won't, never, have not, never before, never, have not yet, never before, by no means, don't, not at all, not at all, forbid, do not, decline, eradicate, have not	−1
Degree Adverb Dictionary	One hundred percent, multiply, unbearable, incredible . . .	2
	however, extremely, miserably, especially, greatly, really/indeed . . .	1.5
	it does not matter, more, more, much . . .	1

References

1. Zhao, L.; Li, Y. Crowdfunding in China: Turmoil of global leadership. In *Advances in Crowdfunding*; Palgrave Macmillan: Cham, Switzerland, 2020; pp. 273–296.
2. Snyder, J.; Mathers, A.; Crooks, V.A. Fund my treatment!: A call for ethics-focused social science research into the use of crowdfunding for medical care. *Soc. Sci. Med.* **2016**, *169*, 27–30. [CrossRef] [PubMed]
3. Renwick, M.J.; Mossialos, E. Crowdfunding our health: Economic risks and benefits. *Soc. Sci. Med.* **2017**, *191*, 48–56. [CrossRef] [PubMed]
4. Jin, P. Medical crowdfunding in China: Empirics and ethics. *J. Med. Ethics* **2019**, *45*, 538–544. [CrossRef] [PubMed]
5. Burtch, G.; Chan, J. Investigating the Relationship Between Medical Crowdfunding and Personal Bankruptcy in the United States: Evidence of a Digital Divide. *Manag. Inf. Syst. Q.* **2019**, *43*, 237–262. [CrossRef]
6. Calic, G.; Mosakowski, E. Kicking off social entrepreneurship: How a sustainability orientation influences crowdfunding success. *J. Manag. Studies* **2016**, *53*, 738–767. [CrossRef]
7. Chen, J.; Chen, L.; Chen, J.; Xie, K. Mechanism and policy combination of technical sustainable entrepreneurship crowdfunding in China: A system dynamics analysis. *J. Clean. Prod.* **2018**, *177*, 610–620. [CrossRef]
8. Ellman, M.; Hurkens, S. Optimal crowdfunding design. *J. Econ. Theory* **2019**, *184*, 104939. [CrossRef]
9. Lee, S.; Lehdonvirta, V. New digital safety net or just more 'friendfunding'? Institutional analysis of medical crowdfunding in the United States. *Inf. Commun. Soc.* **2020**, *23*, 1–25. [CrossRef]
10. Li, Y.; Cao, H.; Zhao, T. Factors affecting successful equity crowdfunding. *J. Math. Financ.* **2018**, *8*, 446. [CrossRef]
11. Sherman, A.; Axelrad, H. A quantitative study on crowdfunders' motivations, their sense of meaning and social welfare. *Int. J. Entrep. Behav. Res.* **2021**. ahead-of-print. [CrossRef]
12. Agrawal, A.; Catalini, C.; Goldfarb, A. Some Simple Economics of Crowdfunding. *Innov. Policy Econ.* **2014**, *14*, 63–97. [CrossRef]

13. Ren, J.; Raghupathi, V.; Raghupathi, W. Understanding the dimensions of medical crowdfunding: A visual analytics approach. *J. Med. Internet Res.* **2020**, *22*, e18813. [CrossRef] [PubMed]

14. Zhao, X.; Mao, Y. The Identity Lies in the Words of Crowd-funders: Help-seekers' Identity Construction in Chinese Online Medical Crowd-Funding Discourses. *Health Commun.* **2021**, *36*, 1–8. [CrossRef] [PubMed]

15. Liu, Q.; Wang, L.; Zhou, J.; Wu, W.; Li, Y. Factors Influencing Donation Intention to Personal Medical Crowdfunding Projects Appearing on MSNS. *J. Organ. End User Comput.* **2022**, *34*, 1–26. [CrossRef]

16. Chen, X.; Sun, M.; Wu, D.; Song, X.Y. Information-Sharing Behavior on WeChat Moments: The Role of Anonymity, Familiarity, and Intrinsic Motivation. *Front. Psychol.* **2019**, *10*, 2540. [CrossRef]

17. Berliner, L.S.; Kenworthy, N.J. Producing a worthy illness: Personal crowdfunding amidst financial crisis. *Soc. Sci. Med.* **2017**, *187*, 233–242. [CrossRef]

18. Gupta, V.; Lehal, G.S. A survey of text mining techniques and applications. *J. Emerg. Technol. Web Intell.* **2009**, *1*, 60–76. [CrossRef]

19. Hearst, M. What Is Text Mining. SIMS UC Berkeley 2003. Available online: http://www.sims.berkeley.edu/~{}hearst/text.mining.html (accessed on 28 December 2021).

20. Bandaru, S.; Ng, A.H.; Deb, K. Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey. *Expert Syst. Appl.* **2017**, *70*, 139–159. [CrossRef]

21. Cios, K.J.; Pedrycz, W.; Swiniarski, R.W. *Data Mining Methods for Knowledge Discovery*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 8.

22. Miettinen, K. Some methods for nonlinear multi-objective optimization. In Proceedings of the International Conference on Evolutionary Multi-Criterion Optimization, East Lansing, MI, USA, 10–13 March 2019; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–20.

23. Sanchez-Gomez, J.M.; Vega-Rodríguez, M.A.; Pérez, C.J. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowl. Based Syst.* **2018**, *159*, 1–8. [CrossRef]

24. Purdy, J.; Nye, P.; Balakrishnan, P. The impact of communication media on negotiation outcomes. *Int. J. Confl. Manag.* **2000**, *11*, 162–187. [CrossRef]

25. Huang, A.H.; Yen, D.C.; Zhang, X. Exploring the potential effects of emoticons. *Inf. Manag.* **2008**, *45*, 466–473. [CrossRef]

26. Daft, R. Information richness: A new approach to managerial behavior and organization design. *Res. Organ. Behav.* **1984**, *6*, 191–233.

27. Simon, H.A. Theories of bounded rationality. *Decis. Organ.* **1972**, *1*, 161–176.

28. Bawden, D.; Robinson, L. The dark side of information: Overload, anxiety and other paradoxes and pathologies. *J. Inf. Sci.* **2008**, *35*, 180–191. [CrossRef]

29. Park, D.H.; Kim, S. The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electron. Commer. Res. Appl.* **2008**, *7*, 399–410. [CrossRef]

30. Xiao, B.; Benbasat, I. Product-related deception in e-commerce: A theoretical perspective. *Mis Q.* **2011**, *35*, 169–195. [CrossRef]

31. Xu, P.; Chen, L.; Santhanam, R. Will video be the next generation of e-commerce product reviews? Presentation format and the role of product type. *Decis. Support Syst.* **2015**, *73*, 85–96. [CrossRef]

32. Li, N.; Wu, D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis. Support Syst.* **2010**, *48*, 354–368. [CrossRef]

33. Greco, F.; Polli, A. Emotional Text Mining: Customer profiling in brand management. *Int. J. Inf. Manag.* **2019**, *51*, 101934. [CrossRef]

34. Karystianis, G.; Adily, A.; Schofield, P.W.; Greenberg, D.; Jorm, L.; Nenadic, G.; Butler, T. Automated Analysis of Domestic Violence Police Reports to Explore Abuse Types and Victim Injuries: Text Mining Study. *J. Med. Internet Res.* **2019**, *21*, e13067. [CrossRef]

35. Zhang, S.; Wei, Z.; Wang, Y.; Liao, T. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Gener. Comput. Syst.* **2017**, *81*, 395–403. [CrossRef]

36. Zeng, X.Q. Technology Implementation of Chinese Jieba Segmentation Based on Python. *China Comput. Commun.* **2019**, *18*, 38–39. [CrossRef]

37. Xu, Q. A Novel Chinese Text Emotion Computation Model. *Comput. Appl. Softw.* **2011**, *28*, 271–272.

38. Wang, S.M.; Ku, L.W. ANTUSD: A large Chinese sentiment dictionary. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 2697–2702.

39. Song, X.-Y.; Zhao, Y.; Jin, L.-T.; Sun, Y.; Liu, T. Research on the Construction of Sentiment Dictionary Based on Word2vec. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 21–23 December 2018; pp. 1–6.

40. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

41. Yi, J.; Yin, M.; Zhang, Y.; Zhao, X. A novel recommender algorithm using information entropy and secondary-clustering. In Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 8–11 September 2017; pp. 128–132.

42. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]

43. Brookes, B.C. The Shannon model of IR systems. *J. Doc.* **1972**, *28*, 160–162. [CrossRef]

44. Wong, S.K.M.; Yao, Y.Y. An information-theoretic measure of term specificity. *J. Am. Soc. Inf. Sci.* **1992**, *43*, 54–61. [CrossRef]

45. Shi, C.Y.; Xu, C.J.; Yang, X.J. Study of TFIDF algorithm. *J. Comput. Appl.* **2009**, *29*, 167–170.

46. Singh, J.P.; Irani, S.; Rana, N.P.; Dwivedi, Y.K.; Saumya, S.; Roy, P.K. Predicting the "helpfulness" of online consumer reviews. *J. Bus. Res.* **2017**, *70*, 346–355. [CrossRef]

47. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]

48. Quinlan, J.R. Combining instance-based and model-based learning. In Proceedings of the Tenth International Conference on Machine Learning, Amherst, MA, USA, 27–29 July 1993; pp. 236–243.

49. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth and Brooks/Cole: Monterey, CA, USA, 1984.

50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

51. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]

52. Schapire, R.; Freund, Y. A decision-theoretic generalization of on-line learning and an application to boosting. In Proceedings of the Second European Conference on Computational Learning Theory, Barcelona, Spain, 13–15 March 1995; pp. 23–37.

53. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 14, pp. 1137–1145.

54. Kvålseth, T.O. Cautionary note about R 2. *Am. Stat.* **1985**, *39*, 279–285. [CrossRef]

55. i-Research. 2018 China's Health Security Industry Report. Financial Network. Available online: https://www.qschou.com/news/content?media-id=133&media-page=1 (accessed on 28 December 2021).

56. CAFP Research. Online Fundraising in China: A research Report on Third Party Platforms in 2014. Available online: https://chinadevelopmentbrief.cn/wp-content/uploads/2020/04/online-fundraising-inChina-EN-cafp20160229.pdf (accessed on 11 November 2015).

57. Zhang, F.; Xue, B.; Li, Y.; Li, H.; Liu, Q. Effect of Textual Features on the Success of Medical Crowdfunding: Model Development and Econometric Analysis from the Tencent Charity Platform. *J. Med. Internet Res.* **2021**, *23*, e22395. [CrossRef] [PubMed]

58. Blum, A.L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271. [CrossRef]

59. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [CrossRef]

60. Lin, H.; Guo, S. Characteristics, scope and classification of degree adverbs. *J. Shanxi Univ. Philos. Soc. Sci. Ed.* **2003**, *2*, 71–74.

61. Xie, X.; Ge, S.; Hu, F.; Xie, M.; Jiang, N. An improved algorithm for sentiment analysis based on maximum entropy. *Soft Comput.* **2017**, *23*, 599–611. [CrossRef]

62. Bakar, N.M.A.; Tahir, I.M. Applying multiple linear regression and neural network to predict bank performance. *Int. Bus. Res.* **2009**, *2*, 176–183. [CrossRef]

63. Sousa, S.; Martins, F.; Alvimferraz, M.; Pereira, M. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ. Model. Softw.* **2007**, *22*, 97–103. [CrossRef]

64. Liu, S.; Cheng, T.; Wang, H. Effects of attention and reliability on the performance of online medical crowdfunding projects: The moderating role of fundraising goal. *J. Manag. Sci. Eng.* **2020**, *5*, 162–171. [CrossRef]

65. Thapa, N. Being cognizant of the quantity of information: Curvilinear relationship between total-information and funding-success of crowdfunding campaigns. *J. Bus. Ventur. Insights* **2020**, *14*, e00195. [CrossRef]

66. Colton, J.A.; Bower, K.M. Some misconceptions about R2. *Int. Soc. Six Sigma Prof. EXTRAOrdinary Sense* **2002**, *3*, 20–22.

67. Leach, L.F.; Henson, R.K. The use and impact of adjusted R2 effects in published regression research. *Mult. Linear Regres. Viewp.* **2007**, *33*, 1–11.

68. Makina, D. The role of social media in crowdfunding. In Proceedings of the ECSM 2017 4th European Conference on Social Media, Vilnius, Lithuania, 3–4 July 2017; Academic Conferences and Publishing Limited: Cambridge, MA, USA, 2017; p. 186.

69. Kang, L.; Jiang, Q.; Tan, C.-H. Remarkable advocates: An investigation of geographic distance and social capital for crowdfunding. *Inf. Manag.* **2017**, *54*, 336–348. [CrossRef]