The Right Time for Crowd Communication during Campaigns for Sustainable Success of Crowdfunding: Evidence from Kickstarter Platform

Michael J. Ryoba 1,2, Shaojian Qu 1,3,*, Ying Ji 1 and Deqiang Qu 1

1 Department of System Science, Business School, University of Shanghai for Science and Technology, Shanghai 200000, China; michael.ryoba@duce.ac.tz (M.J.R.); jijing@usst.edu.cn (Y.J.); qudeqiang@usst.edu.cn (D.Q.)
2 Department of Mathematics, Physics and Informatics, Dar es Salaam University College of Education, University of Dar es Salaam, P.O. Box 2329 Dar es Salaam, Tanzania
3 Department of Decision, Management Engineering School, Nanjing University of Information Science and Technology, Nanjing 210000, China

* Correspondence: qushaojian@usst.edu.cn

Received: 11 August 2020; Accepted: 11 September 2020; Published: 16 September 2020

Abstract: Only a small percentage of crowdfunding projects succeed in securing funds, the fact of which puts the sustainability of crowdfunding platforms at risk. Researchers have examined the influences of phased aspects of communication, drawn from updates and comments, on success of crowdfunding campaigns, but in most cases they have focused on the combined effects of the aspects. This paper investigated campaign success contribution of various combinations of phased communication aspects from updates and comments, the best of which can help creators to successfully manage campaigns by focusing on the important communication aspects. Metaheuristic and machine learning algorithms were used to search and evaluate the best combination of phased communication aspects for predicting success using Kickstarter dataset. The study found that the number of updates in phase one, the polarity of comments in phase two, readability of updates and polarity of comments in phase three, and the polarity of comments in phase five are the most important communication aspects in predicting campaign success. Moreover, the success prediction accuracy with the aspects identified after phasing is more than the baseline model without phasing. Our findings can help crowdfunding actors to focus on the important communication aspects leading to improved likelihood of success.

Keywords: crowdfunding success; phased campaign features; project updates; comments; machine learning; particle swarm optimization

1. Introduction

Crowdfunding is an alternative way of financing entrepreneurs through an Internet-based platform. Crowdfunding platforms offer entrepreneurs an opportunity to seek funds from the crowd to start their projects. Campaigns for the proposed projects are launched on the crowdfunding platforms and interested people in the crowd pledge funds to support the projects. The main stakeholders in crowdfunding usually involve the creators, the ones who seek funds to start a project; the funders (backers), who pledge funds to support projects; and the crowdfunding platform, which is a medium for creators to meet backers [1]. Upon launching a project campaign, creators specify the required fund amounts and the campaign duration, then backers who become interested pledge money to support the project. The campaign becomes successful once the set deadline is over and the total amount pledged becomes equal to or larger than the goal amount specified. Otherwise, the campaign is considered
The sustainable development of crowdfunding platforms depends on the success of campaigns in securing funds from crowds. The more projects succeed in securing funds through crowdfunding platforms, the platforms continue to grow and become more popular in the field.

There are two mostly used schemes in governing crowdfunding platforms, the ‘all or nothing’ and the ‘keep it all’. In the former scheme, the successful projects receive pledged money when the campaign deadline is due, while the unsuccessful projects receive none. On the other hand, creators in the latter scheme receive pledged money after campaign period is due regardless of whether the goal amount is reached or not. Four categories of crowdfunding exist; the reward-based category in which creators look for financial contributions from the crowd in exchange for a reward in terms of a product or service. The second category is the donation-based crowdfunding where creators raise money for a cause and backers are not given back anything tangible in return. The third category is the lending-based crowdfunding, where backers in this category get the binding commitment from creators to repay back the funds at the prescribed interest rate and repay time following their contributions. Lastly, the equity-based crowdfunding where creators sell a specific amount of equity of the proposed project or company and interested backers will receive some ownership in the project in terms of shares [2,3]. This study focuses on the reward-based crowdfunding category that uses the “all or nothing” scheme.

The number of crowdfunding platforms has been increasing since 2006 when the first kind of Crowdfunding platform, the Sellaband platform that aimed at raising funds to support music projects, was launched. Unfortunately, the success rate of campaigns in most crowdfunding platforms is so small, the fact of which puts sustainable developments of crowdfunding platforms at risk. For example, the Kickstarter crowdfunding platform, an ‘all or nothing’ reward-based platform hosted in the US, shows that out of 471,252 launched project campaigns as of December 2019, only 37.65% were successful. High failure rate of campaigns do not only affect creators, but also the reputation and sustainability growth of crowdfunding platforms is affected [4,5]. To ensure sustainable development of crowdfunding, various strategies can be used by platform owners and creators to ensure projects launched in the platforms succeed. For example, platform owners can assess the progress of launched projects through success indicators and recommend the most promising and popular projects to backers. Creators can focus more attention on the success indicators and do what is necessary to fulfil the necessary requirements to ensure their projects succeed. A good number of researches have been done to understand the factors that contribute to the observed dynamics of crowdfunding projects. Previous researchers have focused on identifying features that contribute to the success of campaigns [2,6–11]. Research by Mollick [2] found, using a linear model, that the chances of success are reduced if the project creators do not offer early updates and that the funding success is influenced by project quality, goal amount, and social network ties. Ryoba et al. [6] identified a subset of features with more influence on the success of campaigns using metaheuristic and machine learning algorithms. Ahmad et al. [12] identified, using optimally weighted Random Forests model, 13 features that have high correlation to the success of campaigns.

Various degrees of asymmetric information may have different effects on the success of crowdfunding campaigns. Researchers found that the role of information asymmetry on the success of crowdfunding projects differ from one crowdfunding category to another. For example, Miglo and Miglo [13] investigated the role of information asymmetry when creators have more information than backers and found that equity-based crowdfunding projects suffer more and that asymmetric information favors high-quality projects in reward-based crowdfunding platforms. Belleflamme et al. [3] found that information asymmetry has positive effects on equity-based crowdfunding while it has negative effects on reward-based platforms. Liang, Hu, and Kiang [14] investigated the correlation between information symmetry and success impact of crowdfunding projects and found that frequent information dissemination may not necessarily have positive effects on success likelihood of crowdfunding campaigns. This work supports the possibility of eliminating information asymmetry and therefore realizes the importance of symmetric information between creators and backers. This is because the
only options for backers to infer the quality of proposed projects is through interpreting information provided in the platforms through comments and updates. Furthermore, backers’ funding decisions are influenced by the project information posted during campaigns [15] and thus, it is important to ensure asymmetric information is eliminated. Project updates and comments are the form of online communications that ensure information symmetry between creators and backers. From the perspective of social influence theory [16,17], online communication may influence projects quality perceptions, comment sentiments, and campaign success. Through updates, creators disclose credible information of the proposed projects to enable backers to evaluate the potentials of the proposed projects. Likewise, through comments, creators become aware of the required demands of the proposed projects and backers assess the potential of the proposed projects.

Some researchers [9–11] have investigated the impacts of phased aspects of communication (quantity, quality, sentiment polarity, etc.) extracted from updates and comments on success prediction, but what is missing is an investigation of importance of individual aspect of communication in each phase of the campaigns. Xu et al. [9] investigated the distribution of updates into three phases (the initial phase, the middle phase, and the final phase) and found that for successful campaigns, more updates were offered in the initial phase. Chen et al. [10] conducted research to investigate the impact of phased features on campaign success. They divided the campaign duration into phases and created a prediction model using phased features. They found that prediction accuracy improved considerably in the early phases of campaigns but improved slightly in other phases. Lai et al. [11] assessed the impact of features accumulated in different weeks of campaigns using machine learning approach. Although the authors in refs. [10,11] have phased the campaign duration, they still assessed the combined impact of all features extracted from comments and updates and considered the cumulative values of such features in the subsequent phases. Authors in refs. [9–11,15] focused on investigating the correlation that exists between update and comment features to the outcome of crowdfunding campaigns without assessing the contribution of each feature. In their approach, it is difficult to capture the impact of individual update and comment features occurring in different phases of campaign duration. Only knowing the correlation between the cumulative values for updates and comments features, and the success likelihood irrespective of their distributions over the campaign period, may not be enough to know the important phase and features [7]. Considering cumulative values and the impact of combined features in campaign phases may hinder awareness of success contribution offered by individual features and values that occur in each campaign phase. One may think that more updates are always better and thus attempts to offer daily updates without putting attention to particular update features and campaign phases to focus on. This study intends to bridge this research gap by investigating the success impact of individual time-dependent features extracted from comments and updates that occurred in a specific time slot. Updates and comments features may not have equal success contribution power, therefore investigating contribution of each feature extracted from updates and comments in each campaign phase can enlighten awareness of important features and phases. Knowing the important phases and aspects of communication can offer effective communication and reduce burden on creators and backers. We use a metaheuristic algorithm and machine learning approach to investigate the impact of various combinations of phased features extracted from updates and comments on campaigns’ success prediction. We assess the influence of textual content and frequency of both updates and comments in various campaign phases on success of project campaigns. We split the campaign duration into phases where comments and updates are assigned based on their respective time of posting. Therefore, the important correlation that may exist between individual time-dependent features extracted from updates and comments occurring in various phase with the outcome of campaigns is investigated in this study. Our research intends to provide answers to the following question: Among the phased campaign features extracted from comments and updates, which features and phases are the most important for the campaigns’ success? We anticipate that the findings of our study can enable crowdfunding stakeholders to focus on improving particular aspects of communication in each campaign phase.
A campaign phase is a portion of the entire project duration in which time-dependent features like updates and comments occur depending on the posted timestamp. Authors in ref. [15] analyzed categories of updates that influence the participation of potential backers and found that not just posting more updates is what matters, but rather the specific content of the update. We intend to consider in this study both the right phase for posting project updates and the update content. We argue that just posting updates all the time during the campaign period is not perfect timing. Instead, continuous posting can cause boredom for potential backers or become noisy to classification models. It is more important for creators to learn about the effect of updates and comments and the right time to offer such updates and motivate comments from backers. In so doing, creators will be able to properly manage their campaign by planning updates and motivation strategies for backers’ participation. Proper management of campaigns will save the creators’ time for other project-related endeavors. Furthermore, managed updates can motivate potential backers to be interested in campaigned projects and thus increase the likelihood of project success. Project updates are important during a crowdfunding campaign as they facilitate the crowd to capture the value, credibility, and legitimacy of projects. The design intent of the updates section in the crowdfunding platform is to keep potential backers informed of a project’s progress. Creators may use the updates section at any time during the campaign period to provide any information about the proposed projects. Comments provide the means for potential backers to communicate their opinions or ask questions to the creators about the proposed projects.

We consider the problem of selecting the best combination of features in various campaign phases as a feature selection problem. Feature selection is a combinatorial optimization problem in which computing the optimal solution is usually computationally intractable, especially for problems that are NP-Hard [18]. Therefore, metaheuristic algorithms are used to compute approximate solutions to combinatorial optimization problems. Metaheuristics are approximate methods designed to solve complex optimization problems. They use high-level strategies from different fields like mathematical and physical sciences, nervous systems, biological evolution, and artificial intelligence for exploring and exploiting the search spaces. Using these strategies results in a structured information, which find a near-optimal solution efficiently by intelligently combining different concepts [19,20]. Inspired by the findings of Ryoba et al. [6] on feature subset selection in the crowdfunding context, we extend their study by exploring more on the updates and comments features. This study investigates the success contribution of a combination of update and comment features offered in various phases of crowdfunding project campaigns. Important phases and their features will give insights to creators on the right time to offer and encourage communication with the crowd. We use the particle swarm optimization (PSO) algorithm and the K-Nearest Neighbor (KNN) classification algorithm to accomplish the campaign phase selection problem. PSO is a metaheuristic algorithm first proposed by Kennedy and Eberhart [21] to address continuous optimization problems. Kennedy and Eberhart later proposed the binary version of PSO [22] to deal with binary optimization problems like the feature selection problem. PSO is designed to mimic the social behavior of birds in a flock or fish in a school by searching in the feature space for the best solution of various optimization problems. During the search for solution, PSO finds the optimal feature combination that solves the optimization problem. PSO has been used to solve feature selection problems and was found to easily converge to the optimal solution [23–25]. In this study, PSO is used to generate various combinations of campaign phase features as candidate solutions while the KNN classification algorithm is used to assess their qualities. The motive for choosing PSO in this study is its efficiency in terms of searching capability and speed that facilitates faster convergence to the optimal solution. PSO has faster convergence than genetic algorithms and other evolutionary algorithms, the quality of which makes it suitable choice for feature selection problem [26,27]. The choice of KNN was based on its performance in classifying datasets with very few parameter tuning requirements as compared to other classifiers like Support Vector Machine (SVM). The performance of KNN depends on fine tuning of one parameter concerning the number of nearest neighbors while SVM depends on fine tuning of two parameters and a kernel function.
choice [28,29]. Performing classification without making assumptions about the dataset used [29] and fewer parameters to tune makes KNN a simple and easy classifier to use compared to others like SVM.

This paper contributes to the crowdfunding community in the following ways. First, the study applies a metaheuristic algorithm for assessing the success contribution power of various aspects of communication features in the campaign phases. To the best of the authors’ awareness, the metaheuristic approach has not been used in a crowdfunding context to identify important campaign phases and communication features. Metaheuristic algorithms are effective methods in searching various combinations of phased features that occur during campaign period. They provide a wide search for all possible combinations of phased features and are thus capable of finding the best combination with high prediction performance for crowdfunding campaigns. Secondly, the study contributes to the growing research on crowdfunding by providing relevant insight on the importance of various aspects of communication extracted from comments and updates, and identifying the best combination of such features in predicting the success of campaigns. Owing to the fact that the phased communication features may not have equal importance in determining the successes of campaigns, identifying the best aspect in each phase of the campaign becomes important. The study thus identifies important phased communication aspects that have more impact on the success of crowdfunding project campaigns. The identified phases and features can help crowdfunding stakeholders to focus their attention on the most important phases and aspects of communication that can help in mitigating information asymmetry.

The rest of this paper is organized as follows. Section 2 reviews the literature related to the current study. Section 3 describes the dataset used in this study. Section 4 presents the methods used in this study, which include the feature extraction and processing methods, and the descriptions of PSO and KNN. Section 5 details the experimental settings and the metrics used for evaluation. Section 6 presents empirical results obtained in our study and discussion of findings. Lastly, Section 7 provides the conclusions.

2. Related Work

The contexts of sustainability and crowdfunding have been given attention from different perspectives in the literature. Petruzzelli et al. [30] investigated the role crowdfunding plays in sustainability-oriented projects focusing on campaign design and management, and crowds’ attention triggering in such projects. They argued that effective communication and interaction between creators and backers is important in sustainability-oriented projects as most of these projects offer intangible outputs that might not interest backers. Vismara in ref. [31] analyzed the attractiveness of investors for sustainable-oriented projects in equity crowdfunding. Using regression models, Vismara found that launching sustainability-oriented projects do not increase the chance of success or of engaging professional investors, but rather, attract larger number of crowd who are not professional investors. Cumming et al. [32] used regression analysis to examine the role played by various factors in promoting clean technology projects. They found that the use of soft information in project descriptions, videos, and images to alleviate information asymmetry in clean technology projects has a greater impact on campaigns’ success. Bento et al. [33] used regression analysis to gain insight into the role of project characteristics in influencing the success funding for sustainable-oriented projects and the survival of those projects after the campaign period. Videos, updates, and rewards are among the project characteristics investigated in ref. [33]. The study by Chan et al. [34] researched the association between the language used in communication text and the contribution behavior of backers. Using a regression model, they examined the effects of verbal cues such as money saliency and the intention of sustainability on the number of backers and amount raised. The authors found that as the use of money saliency in project descriptions increases, the most negative effects can be observed on both the amount raised and the number of backers. However, they found that this is different for sustainability-oriented projects as sustainability contexts mitigate the negative effects of money saliency. Bento et al. [35] investigated the impact of the risk profile of crowdfunded clean technology projects on investors’
returns. They found that technology risk decreases the excess of return of projects. Hörisch [36] analyzed factors influencing sustainable projects to succeed in financing and marketing using linear and logistic regression models. Updates made during campaigns are among the identified features with positive influence for sustainable projects to meet their funding targets. However, the impact of individual campaign phased features on crowdfunding project campaigns remains uncertain.

Another line of research in sustainability and crowdfunding focused on factors influencing sustainable growth of crowdfunding platforms and sustainable financing of projects [4,14]. Fernandez-Blanco et al. [4] investigated the key factors to the success of campaigns for sustainable crowdfunding platforms. Fernandez-Blanco et al. [4], using a data mining approach, found that comments and updates are among the factors that have a positive correlation with campaign success. Liang et al. [14] assessed the impact of information descriptions provided through updates, comments, and project descriptions on success likelihood of campaigns for sustainable financing of proposed projects. From the perspective of information communication, information asymmetry, and signaling theory, Liang et al. [14] identified the quality, attitude, and quantity aspects as used in the project communication descriptions and assessed their impacts on campaign success using binary logistic regression. They found that updates and comments were among influencers for campaign success while readability has negative impact on campaign success. For platforms and projects to guarantee sustainable financing and growth, much emphasis should be put on the effective communication through comment and updates occurring in most important campaign phases.

Information asymmetry occurs when one part lacks information concerning the potential and quality of what is proposed by another part [37,38]. Information asymmetry is the major cause of failures in traditional credit markets [14,39]. Similarly, in a crowdfunding context, information asymmetry poses challenges to backers and creators, and has a negative influence on the success of campaigns [14,40]. Creators may know much more about the quality of the project they are proposing compared potential backers, while backers may be unaware about the credibility of creators in producing and delivering the promised products or services [2,40,41]. To reduce information asymmetry between parties, the signaling theory suggests that the informed party can send signals that can help the less-informed party to be informed [42]. Uncertainty and information asymmetry are common phenomenon in crowdfunding context where proposed projects are usually not yet in market in a finished state [3]. Courtney et al. [38] used information economics theory to investigate the success impact of signals obtained from project descriptions and updates, and backers’ sentiments expressed through comments. Using logit regression they found that signals from projects and sentiments reflected from comments help to mitigate information asymmetry concerning project quality and creators’ credibility and have positive impact on crowdfunding success. Based on the perspective of ecosystem, Kang and Kim [43] investigated the impact of social-technical ecosystem on the survival and evolution of crowdfunding participants. From an ecological perspective, the evolution process occurs, which determines the condition for survival of participants. The evolution process is accelerated by the dynamics and changes in social and human factors influenced by the participants’ relationships and interactions [43,44]. The sustainability of crowdfunding platforms and projects are achieved when crowdfunding evolution advances in a positive direction. Using regression analysis, Kang and Kim [43] found that interactions in terms of feedback and feed-forward between creators and backers have positive impact on the campaign success. Belleflamme et al. [3] investigated the role of uncertainty and information asymmetry on the design of crowdfunding projects. They found that under information asymmetry, the entrepreneurs prefer a profit-sharing scheme than pre-order in determining the financing option of their projects. Davies and Giovannetti [45] developed a conceptual framework to investigate the role played by crowdfunding platforms in facilitating signaling activities as the means to alleviate the negative effects of asymmetric information. They considered, in their model, the signaling originated from both creators and backers in terms of social capital, reciprocity, and experiences. Using a logit regression model, the authors in ref. [45] found that the assessed signaling has a positive impact on the success of project campaigns.
The influence of backers’ decisions through the quantity, contents, usage patterns, language, and sentiments is investigated in refs. [15,46,47] and found to have positive impacts on campaign success. The way backers communicate and interpret information helps in spreading word-of-mouth signals to potential backers, which influences funding decisions. Backers can make informed funding decisions based on projects’ legitimacy and credibility established through quality signals sent by creators via updates [15,48]. Project updates are the most important avenues for creators to express the qualities of the proposed projects. High-quality and credible projects trigger the crowd’s acceptance and participation which can be observed through comment sentiments and funding rates [2,15,49]. The comments section in crowdfunding platforms is used as the channel for backers to interact with creators and other backers. Individuals who are interested in a proposed project tend to seek opinions from others through posted comments. Active interactions between creators and backers during the campaign period via updates and comments create information symmetry between the two parties. The quality of updates can tell the preparedness of creators in accomplishing the proposed projects. Projects with more updates help in reducing uncertainty to backers and therefore, increase the chances of being funded. On the other hand, more negative comments create bad impression to potential backers and negatively affect the funding process for the projects. Block et al. [15] used regression analysis method and noted that the updates that used easy language and that were posted earlier during campaigns were more influential.

In investigating the distribution of project updates, Xu et al. [9] conducted a study that divided the campaign duration into three equal phases (the initial phase, the middle phase, and the final phase). They assessed the distribution of updates in different phases and noted that more updates were on the initial phase. Wang et al. [46] investigated the impact of comment features on success prediction and noted that comment sentiment and quantity were among the features that have influences on the outcome of campaigns. The study by Lee et al. [50] confirmed that text data from updates and comments have much influence on success prediction of campaigns. Their model created with only text data from updates and comments was able to predict campaign success with an accuracy of 85–91%. Comments were found to have more impact after 5–10 days while updates’ impact was observed from day 20–30. Niemand et al. [7] studied the non-linear effects of project updates on the success of the crowdfunding campaigns. They used optimum stimulation level theory and marginal utility theory to explain the observed non-linearity of project features. Authors in ref. [7] concluded that the number of project updates had decreased effects on the crowdfunding success with a maximum of five updates per project as the saturation point. Their findings suggest that cumulative update features might have less effect on the success likelihood of the crowdfunding projects and thus call for more research on phase-wise updates.

The effects of phased time-dependent features on success prediction have been investigated in literature using machine learning approach. Division of campaign duration into phases makes it possible to assess the correlation between phased features and the success of the campaigns. Chen et al. [10], in their study, created a series of prediction models using random forest classifier. They used a combination of time-dependent and static features of project campaigns to perform predictions in different phases of the campaign. In the crowdfunding platform, the campaign duration is set in terms of days. The fractions of campaign duration can be considered as campaign phases. Chen et al. [10] divided the campaign duration into seven phases and assigned time-dependent features in each phase to contain accumulated values of previous phases. Their eight classification models show that prediction accuracy improved considerably in the early phases of campaigns but remained comparable or improved slightly in other phases. The decision tree approach was used by Rao et al. [51] to analyze the correlation between the pledged amount, the time-series data, and the likelihood of the campaign success. They divided the campaign duration into phases of 5% of length of the respective campaign duration. Rao et al. [51] then analyzed phases of the campaign duration that are more or less predictive by creating models trained by predictors with time-series data before the respective phase.
They found that pledged amounts that occurred in the initial 10% and between 40–60% of the duration period had the strongest impact on the campaign success.

Lai et al. [11] conducted a study to examine the influence of comments and updates on funding success across different campaign durations using machine learning (eXtreme Gradient Boosting) and text-mining (lexicon-based sentiment analysis) techniques. The authors divided the campaign duration into three phases; the first week, the first two weeks, and the entire campaign duration. The posted comments and updates were then assigned into these phases depending on the posting time. Authors in ref. [11] used the boosting models to determine whether the features of their interest were really important or not. The authors then created a predictive model using all their features without performing selection on the most important ones. They found that textual features (readability, sentimental words, and language used) from comments and updates that occurred in the first week are better than those in the second week in success predicting task. Ryoba et al. [6] used WOA metaheuristic algorithm and KNN to assess success prediction of various feature subsets. The authors identified the best subset of features that had an accuracy of 90.28% in predicting the outcome of crowdfunding projects. The number of comments and updates were among the features in their identified subset, but only cumulative values for the number of updates and comments covering the entire campaign period were considered in their study.

The literature discussed in this section demonstrate the contribution power of interaction between backers and creators to the success of campaigns in crowdfunding platforms. Their results provide insights on the contribution power of comments and updates posted during the campaign duration to the success of campaigns. This study assesses the contribution power of comment and update features occurring in various campaign phases. We extend the work of Ryoba et al. [6] by exploring more on the update and comment features by considering the time such updates and comments were posted. Our focus is to evaluate the predictive power of features in various combinations of campaign phases when combined with control variable features identified by Ryoba et al. [6]. Identifying the most important campaign phases and their features can help creators in performing necessary efforts to increase the success likelihood of campaigns. Different from refs. [10,11,51], this study uses a metaheuristic algorithm and machine learning to assess the correlation of features in various combinations of campaign phases towards success of crowdfunding project campaigns. Metaheuristic algorithms are more effective methods for searching optimal features combination than exhaustive and random search methods [52–54]. We used metaheuristic algorithm to select the best combination of campaign phase features and used them to create a prediction model. Unlike the approach used by Chen et al. [10], we used the best set of phased campaign features in creating a single classification model and that our phased features include features occurring in respective phases. Although Lai et al. [11] analyzed the impact of features on phases based on accumulation of features in weeks, they did not analyze the impact of other set of weeks. For example, the impact of the first week and the third week, the second week and the third week, and the first week and the third week are not reported. Furthermore, in their sentiment analysis, authors in ref. [11] did not consider the sentiment polarity aggregates of comments for each project, they only considered the sentimental words found in comments. The sentimental polarity of all comments is important for proper evaluation of sentiments of all backers who commented about the project and, thus, gives proper signal that may impact the funding decisions of future backers. In this work, we consider all possible combinations of phases and the aggregate sentiment polarity of comments in assessing their impact in success prediction. Similar to refs. [9–11,51], we divide campaign duration into phases with features assigned based on their timestamp and use machine learning techniques to create a prediction model. We analyze textual features in all updates and comments assigned in each phase using a semantic analysis similar to the study in refs. [9,11].
3. Dataset Description

This study uses Kickstarter dataset for project campaigns launched between December, 2018 and June, 2019. A significant number of researchers [1,10,12,55,56] have used datasets from the Kickstarter crowdfunding platform, which is the most famous, oldest, and largest in the field. The statistics from Kickstarter platform show that there were approximately 17 million backers as of December, 2019 who pledged around 4.8 billion dollars to support proposed projects (https://www.kickstarter.com/help/stats). Kickstarter crowdfunding platform is revealed from these facts to be the most influential platform and, thus, using the Kickstarter dataset means the findings obtained are reliable and trustworthy. The Kickstarter platform provides public access to various information related to project characteristics and performances for all projects, be it active, successful, suspended, cancelled, or unsuccessful. Project information, which is accessible publicly, includes among others: The project description, reward description, and the updates on the projects as well as the number of such updates. Other publicly accessible information includes: The comments posted by creators and backers, the required fund amount, and the campaign duration. This study used the following publicly accessible project features: Comments, updates, comments timestamp, updates timestamp, creators’ backed projects, creators’ launched projects, pledge descriptions, and the number of videos used. Others are: The reward descriptions, goal amounts, campaign duration, campaign start time, campaign end time, the campaign status, and number of reward choices. The choice to include these project features in our study is based on their success prediction power as reported by Ryoba et al. [6] and on the intention of this study i.e., assessing the contribution of phase-wise features. We used pre-scraped dataset from a scraper robot (https://webrobots.io/kickstarter-datasets) to obtain some of the project features of our interest. The scraper robot crawls the Kickstarter platform for data extraction on monthly bases.

Furthermore, we used our own scraper to crawl the Kickstarter platform for extraction of other features of interest to this study that were missing in the pre-scraped dataset. The missing features of interest includes: Project description, number of updates, reward description, number of videos, comments timestamp, number of comments, updates timestamp, and number of projects that the creators created and backed. For the posted project updates, we were not able to get the updates that creators posted privately for their backers; only publicly accessible updates and their timestamps were then collected. For all projects in the collected dataset, we converted the goal amounts to the same currency (US Dollars) to ease the assessment. Projects with the status of ‘live’, ‘suspended’, or ‘canceled’ were removed and only the ‘successful’ or ‘failed’ projects were retained. Since the aim of this study is to assess the impact of features extracted from updates and comments, we filtered the dataset to contain projects with at least two publicly accessible updates and comments. To deal with unbalanced data in terms of the number of success and fail campaigns in the collected dataset, we performed random under-sampling technique similar to the approach used in ref. [6,12]. Thus, only 15,270 projects remained for further processing as shown in Table 1 with other statistics of the dataset included.

| Table 1. Project dataset statistics. |
|--------------------------------------|
| Total Projects                       | 15,270 |
| Successful Projects                  | 7635   |
| Failed Projects                      | 7635   |

4. Methods

4.1. Data Preprocessing and Feature Extraction

Various features from Kickstarter projects have been considered in different studies when examining success prediction of project campaigns [6,9–11]. Features vary in success prediction power and therefore various feature selection processes have been used to filter the most important features from the less important ones [6,10,11,51]. Inspired by the findings on feature subset selection by
Ryoba et al. [6], this study considered features obtained by their study together with an addition of features related to updates and comments as the baseline features. Since we intend to assess the effects of phase-wise textual features extracted from comments and updates, we incorporate additional features to our baseline features to include sentiment polarity of comments and readability index of updates. In the baseline features, the campaign status is used as the dependent variable which allows us to assess the impact of update and comment features on the outcome of this variable. Preprocessing is performed on the dataset to obtain some of the baseline features. Stop words, punctuations, and numbers were removed in all textual features in the dataset before extracting the features of interest. Moreover, we used text-mining techniques detailed in Section 4.1.1 to extract features of interest from update and comment texts. The baseline features therefore include: Project description length, update fog-indexes, number of reward words, comments polarity, number of videos, number of projects backed by creators, funding goal amounts, number of projects created by creators, number of comments, number of reward levels, and number of updates. The dataset was then filtered to contain only baseline features as described in Table 2. To better understand the contribution of phased features, we considered static features from our baseline features as our control variable features. Considering that project campaign is conducted for a given duration, we divided campaign time into phases with 20% of the campaign duration as one interval and assigned dynamic features on each phase depending on its timestamp. In this case, we have five phases for each project campaign each phase having four features extracted from comments and updates posted during that time. Since most of the Kickstarter projects have an average of 30- or 60-day duration, the 20% means an interval of 6 days for a project of 30 days duration and 12 days for 60-day duration. The dynamic features assessed in this study are features extracted from project updates and comments posted by creators and backers. The aim is to examine the success contribution of features in a combination of the campaign phases.

| Project Features | Feature Status | Description |
|------------------|---------------|-------------|
| Goal             | Static        | An amount set by a creator required for starting a project |
| project_words    | Static        | The number of words used in the project description |
| reward_words     | Static        | The number of words used in the reward description |
| number_video     | Static        | The number of videos used in a project description |
| pledge_levels    | Static        | The number of reward levels used in the reward description |
| campaign_status  | Static        | The status of a project campaign used as a dependent variable (class label) |
| Created          | Static        | The number of projects created by creators |
| Backed           | Static        | The number of projects backed by creators |
| comments_polarity| Dynamic       | The sentiment expressed in comments, the positive and negative polarity |
| num_comments     | Dynamic       | The number of comments posted by backers or creators |
| updates_fog-index| Dynamic       | The readability index of updates |
| num_updates      | Dynamic       | The number of updates posted by creators |

4.1.1. Text-Mining Methods

This study extracted text-mining features from project updates and comments assigned in each campaign phase. We performed sentiment analysis on each comment to extract sentiment words and determined the polarity of comment by computing the difference between the positive and negative word counts. Sentiment analysis refers to the process of discovering opinion expressed in texts. The polarity of comments may affect potential backers’ intention to fund projects as the decision of some people is affected more by sentiment orientation. Wang et al. [57] assessed the impact of sentiment
orientation on the success of campaigns and found out that sentiment and textual quality features derived from project comments further boost the prediction accuracy. To perform sentiment analysis on comments, we used lexicon-based sentiment analysis [58], which performs the analysis by extracting opinion words. For example, given the sample comments: “Thanks for the clarification concerning international supporters in the updates. Amazing project! Congratulations for a great idea! I’m slightly sad to see USA only on most pledges”. The opinion mined by lexicon-based sentiment analysis on the comments results in 4 positive and 1 negative words. The computation of polarity difference results in +3 and therefore the general polarity of comments is positive with a value of 3. For update text, we extracted features that signal the quality of project. The quality signal from updates determines the preparedness and confidence of creators in the proposed project, and therefore, high-quality signal attracts more people while the low-quality one will distract people. To assess the quality of updates, we extracted the readability index using the gunning-fog-index and the number of words used in the updates.

4.2. Feature Selection Methods

Candidate solutions to the campaign phase selection problem was represented using ‘1’ and ‘0’ binary values similar to the representation by Ryoba et al. [6]. The value ‘1’ means the phase feature was selected and ‘0’ means it was not selected. The one-dimensional vector was used to represent each candidate solution where the value in each vector component is either ‘0’ or ‘1’ and the length of a vector is the total number of features in all phases. Binary PSO was then used for searching various combinations of phased features as candidate solutions to the problem. Since the intention was to find the minimal number of phased features and higher classification accuracy, we used the fitness function given in Equation (1) that considers the two objectives [6,59]. The KNN classification error rate $\gamma_s(R)$ of a subset of phased features in Equation (1) is computed using Equation (6).

$$Fitness = \alpha(1 - \gamma_s(R)) + \beta \frac{|S|}{|N|}$$

(1)

In Equation (1), $|S|$ is the total number of selected features and $|N|$ is the sum of features in all campaign phases in the dataset. $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$ are parameters used to achieve a balance of the two important objectives, the classification accuracy and the length of a subset.

Particle Swarm Optimization (PSO) Algorithm

In this work, binary PSO is used to search for the candidate solutions to the campaign phase selection problem. The target was to search a combination with minimal number of phased features that predicts crowdfunding projects with high accuracy. PSO is a metaheuristic algorithm that was proposed by Kennedy and Eberhart [21] and later the same authors proposed its binary version to solve the binary optimization problems [22]. PSO was designed to mimic the social behavior of the swarm of birds or fishes when searching for optimal path to the food source. PSO has been used with much success in a wide range of problems ranging from function optimization, fuzzy system control, feature selection, and artificial neural network [60,61]. PSO has been proved to converge faster to the optimal solution than Genetic Algorithm (GA) and other evolutionary-based metaheuristic algorithms [27,61]. This fact makes PSO a suitable algorithm for searching the feature space. Moreover, PSO is simple to implement and computationally feasible compared to other well-known algorithms like GA in solving optimization problems [62]. PSO consists of a set of particles that represent birds or fishes. Each particle has to move in a search space with a certain velocity to obtain the best path towards the location of food. The position vector of a particle in a binary PSO contains binary values of 0 or 1, which suit the binary optimization problems. Based on the nature of our phased feature selection, this study adopted the binary version of PSO. The particles in binary PSO are used to represent the candidate solutions of the phased campaign feature selection problem whose features need to be evaluated in terms of their prediction power.
In PSO, the movement of the swarm towards the position of the best particle is achieved when particles adjust their positions by changing velocity. Each particle changes velocity by learning from its own experience and from the experience of other particles. The best encountered position for each particle and among all particles are recorded as personal best (pBest) and global best (gBest), respectively. A particle changes velocity and position using Equations (2) and (3), respectively.

\[ V_{t+1}^i = w \cdot V_t^i + c_1 \cdot r_1 \cdot (pBest_i - X_t^i) + c_2 \cdot r_2 \cdot (gBest - X_t^i) \]  
\[ X_{t+1}^i = X_t^i + V_{t+1}^i \]  

In Equations (2) and (3), \( t \) represents the iteration number, and \( X_t^i \) and \( X_{t+1}^i \) denote, respectively, the positions of particle \( i \) in iterations \( t \) and \( t + 1 \). The variable \( w \) represents the inertia weight that controls the exploration and exploitation phases of the algorithm and \( c_1 \) and \( c_2 \) are the learning constants. The variables \( r_1 \) and \( r_2 \) are uniformly distributed numbers in the range \([0, 1]\), \( pBest_i \) and \( gBest \) are the best position of particle \( i \) and the global best position among all particles, respectively. \( V_t^i \) and \( V_{t+1}^i \) are velocities at iterations \( t \) and \( t + 1 \), which are constrained by the predefined values usually in range of \([-V_{\text{max}}, V_{\text{max}}]\).

In binary PSO, the particle positions, personal best position and global best position, can only have binary values of 0 or 1. The particle’s velocity is updated using the same approach expressed in Equation (2) as used in the standard PSO. The binary values for the particle’s position are obtained using Equation (4) by applying the sigmoid function in Equation (5) on the particle’s velocity.

\[ X_{t+1}^i = \begin{cases} 1 & \text{if } rand < f(V_{t+1}^i) \\ 0 & \text{otherwise} \end{cases} \]  
\[ f(V_{t+1}^i) = \frac{1}{1 + e^{-V_{t+1}^i}} \]  

In Equation (4), \( rand \) is a random number in the range \([0, 1]\).

Figure 1 is the pseudocode for binary PSO used in this work. In the first step of the algorithm, initialization of parameters is done. Among them is generation of a random population of the swarm. Then, the fitness of each particle is computed using Equation (1) and the global best position, which is the best solution so far, is identified. The algorithm then performs a number of iterations while searching for the other global best position. After a set number of iterations, the algorithm returns the value for the overall global best particle position in all iterations and its fitness as the solution to the campaign phase selection problem.
1. **Input**: Population size $n$, maximum number of iterations $MaxIt$, the minimum and maximum inertia weight values $\omega_{\text{Min}}$ and $\omega_{\text{Max}}$, the learn factors $c_1$ and $c_2$, the maximum velocity $V_{\text{max}}$ and the dimension of particles vector $\text{dim}$.

2. **Step 1** Initialization:
   
   3. **Step 1.1** Generate initial population vector $X_i^0$ ($i = 1, 2, ..., n$)
   
   4. **Step 1.2** Initialize personal best and global best position vectors and their scores:
      
      $g_{\text{Best}} \leftarrow \text{Null}; g_{\text{BestScore}} \leftarrow -1$
      
      $p_{\text{Best}}_i$ ($i = 1, 2, ..., n$) $\leftarrow \text{Null}, \quad p_{\text{BestScore}} \leftarrow -1$

   5. **Step 1.3** Generate particle’s initial velocity vector:
      
      $V_i^0$ ($i = 1, 2, ..., n$) $\leftarrow \text{random} (-1, 1)$

   6. **Step 2** Begin cycling and set iteration counter $t = 0$

   7. **Step 2.1** Calculate the fitness value $\text{fitness}_i$ ($i = 1, 2, ..., n$) for each particle using Equation (1)

   8. **Step 2.2** if $\text{fitness}_i > p_{\text{Best}}_i$ update the personal best position vector and personal best position vector score

   9. **Step 2.3** if $\text{fitness}_i > g_{\text{Best}}$ update global best position vector and global best score

   10. **Step 2.4** For each particle in the population:
      
      Update particle velocity using Equation (2)
      
      Update particle position using Equation (4)

   11. **Step 3** Stop criteria: if $t = MaxIt$ stop the algorithm; otherwise set $t \leftarrow t + 1$ and go to **Step 2.1**

12. **Output**: Output the $g_{\text{Best}}, g_{\text{BestScore}}$

**Figure 1.** Pseudocode for a Binary PSO ([25]).

4.3. **Classification Methods**

The prediction of a crowdfunding campaign involves two possible outcomes, success and failure, and as such, the prediction task is regarded as binary classification. While a number of binary classification algorithms exist, we opted to use KNN algorithm in this paper. KNN algorithm is a widely used classifier with high performance on a wide range of real datasets [63]. To avoid overfitting of classification model and to ensure robust results, we used $K$-fold cross-validation technique to divide the dataset into two sets, the training set and the testing set. The training set of the dataset included class labels and were used for training the KNN classification model. On the other hand, the testing set was used to test the KNN model by predicting its hidden class labels. Principally, KNN works by comparing distance of each data instance in the testing set to all data instances in the training set. Then, for each data instance in the testing set, the $K$ nearest neighbors are examined. The class labels of the majority from the $K$ nearest neighbors is thus taken to be the predicted class label of the respective data instance in the testing set [29,64]. The predicted class of each data instance in the testing set is then
compared to its real class label (its hidden class values). The prediction accuracy is then computed by considering the number of times the model is right in prediction.

The outcome of crowdfunding projects in the dataset was used as a class label, which is assigned a ‘1’ for successful projects and a ‘−1’ for failed projects. To avoid bias of the model in the learning process, we used random under-sampling technique to maintain a one-to-one ratio for the two classes in the dataset. Using K-fold cross-validation, we divided our dataset into K equal portions similar to ref. [65] in which (K−1)−folds of the dataset are combined to form the data portion used for model training and the remaining one fold is used as the data portion for testing the model. This process of training and testing the model is repeated K times while interchanging, in each time, the data portions for model training and testing. The average result for the classification error rate \( \gamma_s(R) \) is then computed by Equation (6) similar to refs. [6,28,59].

\[
\gamma_s(R) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} f(c(x_n), y_n)
\]  

In Equation (6), \( K \) represents the total number of iterations, \( N \) is the number of data instances in the testing set, \( c(x_n) \) denotes the predicted class of the \( n^{th} \) instance in the testing set, \( y_n \) represents the \( n^{th} \) instance actual class label, and \( f(c(x_n), y_n) \) denotes a function that is evaluated as to ‘1’ if \( c(x_n) \neq y_n \), and ‘0’ otherwise.

5. Experiments

5.1. Algorithm Implementation

We implemented PSO using Python programming language and used Python standard libraries (sklearn developed by Pedregosa et al. [66], NumPy [67], Pandas [68], and NLTK [69]) to carry out the experiments. The NLTK library was used to remove all stop words, punctuations, and numbers as well as to tokenize all comment texts. Then, from the opinion lexicon of the NLTK module, we determined the difference between the two polarity counts of respective comments, positive and negative counts. A KNN classification algorithm from the sklearn package was used for the classification task and for partly assessing the quality of all the intermediate solutions obtained using PSO. Two experiments were conducted: The first experiment used PSO to perform the selection of the best combination of phased communication features extracted from updates and comments. To perform the selection task, 10 runs of experiment one were conducted whereby in each run, PSO iterated 100 times while searching for different combinations of phased communication features. The aim was to select among the evaluated combinations, the one with the largest fitness value in each run of experiment. The best selected combination of phased aspects of communication was then combined with control variable to perform success predictions of crowdfunding campaigns. In the second experiment, classification was performed using our baseline features discussed in Section 4.1 and presented in Table 2. Dynamic features in the second experiment were not phased, rather they contained their cumulative values that covered the entire campaign period. Thus, phased feature selection was not performed and in that case PSO was not used at all. The same classifier (KNN) was used to perform classification task in the second experiment. To prevent model learning bias, we performed normalization of all features in the dataset to the scale of \([0, 1]\) using MinMaxScaler in preprocessing module of the sklearn package.

5.2. Parameter Settings

The parameters for all experiments were set as follows. For K-fold cross-validation, we set a K to be 10, similar to the studies by Mafarja and Mirjalili [59] and Ryoba et al. [6]. For the 10-fold cross-validation used, 9-fold was used for training the model and the remaining fold was used for testing it. The process was repeated 10 times while interchanging the parts of the dataset used in the training and testing. We set the binary PSO parameters as follows. The maximum number of iterations
was 100, the population size was 50, the inertia weight \(w\) was set to vary from 0.2 to 0.9 as iterations proceeded, and the learning parameters \(c_1\) and \(c_2\) were both set to be equal to 2. One or more of these settings were also used in refs [70,71]. Experiment one was run 10 times. Then, the average result was taken as the final solution to the campaign phase selection task. KNN classification algorithm with Euclidean distance metric was used in the two experiments. After experimenting with different values of \(K\), we found 25 as the best. All experiments were performed using an Intel Core i5 machine with 4GB of RAM and 2.4 GHz CPU.

5.3. Evaluation Metrics

Confusion matrix is the widely used tool in classification field to show prediction efficiency of classification models. A confusion matrix is a table that describes the performance achieved by a classifier on the test dataset instances. The rows and columns of the table represent, respectively, the actual and predicted class of the test dataset instances (or vice versa). All other evaluation metrics used in classification field (accuracy, precision, recall, and F-score) are derived from a confusion matrix. The confusion matrix for binary classification problems, like crowdfunding prediction problem, usually contains four values; the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TN and TP indicate the correct frequency of the model in classifying data instances as negative and positive, respectively; whereas the FN and FP show the incorrect frequency of the model in classifying data instances as negative and positive, respectively [29].

The evaluation metrics derived from confusion matrix and used in this work to assess the models’ performance include the accuracy, F-score, recall, and precision. Accuracy is among the most-used metrics in assessing the performance of a classifier and its computation is performed using Equation (7).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}
\]

Precision of a classifier is defined as the fraction of relevant predictions from the predicted group in a classification task. For positive instances, precision is computed using Equation (8).

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{8}
\]

Recall of a classifier shows the fraction of relevant predictions made from the actual relevant instances, and for positive instances it is computed using Equation (9).

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{9}
\]

F-score (or F-Measure) of a classifier is computed by finding the harmonic mean of the precision and recall values of a classifier. It is calculated by using Equation (10).

\[
F \text{- score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}
\]

6. Results and Discussion

Various combinations of campaign phases composed of different communication features that predict campaign success with different accuracies were evaluated in experiment one. The best combination of phased communication features extracted from comments and updates was then selected. The best selected combination represents the most influential phases and communication features in success prediction of Kickstarter crowdfunding campaigns. Table 3 shows the best chosen phases and their communication features with the largest fitness value for each experimental run. On average, five phased communication features were obtained after 10 runs of selection experiments and were distributed in various campaign phases. The importance of features from the identified
combination of phases was then assessed. The frequency of repetition of a feature in different runs of experiments determines its importance in success prediction task. Table 3 also shows the frequency of repetition of each identified feature.

Table 3. Selected phased features and their frequencies of reputation.

| Run | Number of Features | Phase 1 (20% CD) | Phase 2 (20–40% CD) | Phase 3 (40–60% CD) | Phase 4 (60–80% CD) | Phase 5 (80–100% CD) |
|-----|--------------------|------------------|---------------------|---------------------|---------------------|---------------------|
| 1   | 3                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 2   | 4                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 3   | 4                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 4   | 4                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 5   | 5                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 6   | 5                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 7   | 4                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 8   | 6                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 9   | 7                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |
| 10  | 6                  | ✓                | ✓                   | ✓                   | ✓                   | ✓                   |

Frequency of repetition: 0 0 9 0 0 7 0 1 0 10 0 5 2 1 3 4 0 5 0 1

Key: F1—Number of comments; F2—Sentiment polarity of comments; F3—Number of updates; F4—Readability index of updates; CD—Campaign Duration.

From the average number of features obtained in this study, we consider the first five features from various phases with the highest frequency of appearance in the experiment as the best identified feature combination (see Table 3). The identified most important features for success prediction are found in different phases. It should be noted that the identified phase becomes important in the context determined by its chosen feature(s) and not all features in that phase. The most important phases in this case are phase one, two, three, and five. The important features in these identified phases are as follows:

The number of updates in phase one; the polarity of comments posted in phase two; the polarity of comments and readability of updates posted in phase three; and the polarity of comments occurring in phase five. The identified features signify the important aspect of communication required in the respective phase to lessen information asymmetry between backers and creators. The number of updates in phase one indicates that the communication in phase one has to focus on the quantity of updates. The more updates provided in this phase allows potential backers to assess, in the earliest stage, the potentials of the proposed projects using more potential information and help them to make informed decision. As for the polarity of comments in phase two, three, and five, it means that in these phases, the communication that has a focus on the sentiment expressed via comments is much more important. The sentiments from comments aid in spreading word-of-mouth effects concerning the potential ability of the proposed projects to potential backers. Finally, the readability of updates in phase three signifies that the quality of updates matters a lot in this phase. The quality of updates allows potential backers to capture the credibility of the proposed projects. The asymmetric information is mitigated once the focus is put in the identified aspects of communication in the respective phases for the success of campaign and sustainability of project financing and growth of crowdfunding platforms.

Our method is able to assess, in each phase, the individual contribution of update and comment features (the number of comments, the sentiment polarity of comments, the readability index of updates, and the number of updates) and select the most important among the four features in each phase. However, the individual importance of comment and update features in each campaign phase have not been assessed in literature. In general, the important phases on success prediction obtained in our study are in agreement with most of findings in literature. Since most of the research works in the
literature considered the initial days, subsequent/middle days, and last days in presenting their findings, we consider from our settings the initial days to be days that fall in phase one and phase two, middle days/subsequent days to be days that fall in phase three, and last days to be days that fall in phase five for comparison purposes. References [2,10,11,72] found that comments and/or updates occurring in the initial days of the campaign had more impact on success prediction. Lai et al. [11] found that all features extracted from comments and updates (the readability, sentimental words, and language used) are influential for success prediction in the initial stage of campaigns. Mollick [2] found that updates, without considering any specific feature, which are posted soon after launching campaigns, determine the success. Chen et al. [10] and Gera and Kaur [72] both found that dynamic features that include updates and comments (only the number of updates and comments were assessed) are more influential in early phases of campaigns. We found that the important features in the initial days are only the number of updates and the polarity of comments. In the initial days, the number of updates signifies that it is the quantity of updates that matters in this particular stage. The quantity of updates can help in mitigating asymmetric information between creators and backers, which lead to improving backers understanding on the potentials of the proposed projects. The sentiment polarity of comments occurring in the initial days of campaigns help to express the reaction of crowds in the earliest stage where the positive reactions assist in attracting more backers to fund the proposed projects while the negative reactions destruct potential backers in offering their support. Chen et al. [10] found that the dynamic features improved the accuracy as campaigns progressed to the subsequent phases, but found it fixed or improved slightly in some phases. Likewise, Lai et al. [11] found that the comments and updates in the subsequent phases contributed less to the success prediction. Although we differ with Chen et al. [10] and Lai et al. [11] on the update and comment features assessed as detailed before, we also found that the middle phase is important for success prediction for updates that emphasized in quality (readability) and comments in terms of sentiment polarity. From our findings in the middle phase, it can be concluded that the phase is used by the potential backers in assessing what others say about the proposed projects and how the creators are prepared to fulfil the promised targets. Lastly, we found that last phase is important in the polarity of comments. This shows that the campaign’s last minute is used by the potential backers to assess reviews from others before making up their minds.

We conducted another experiment to aid in evaluating the impact of our phased model on the prediction performance. In this second experiment, we used our baseline features discussed in Section 4.1 to perform the prediction task. We refer in this work to the model created in the second experiment as the baseline model, and the model created in experiment one as phased model. Our identified phases and their features created a predictive model (phased model) that predicts the outcome of crowdfunding projects with an accuracy of 91.8%. The obtained accuracy is an increase of 0.9% from the baseline model in classifying projects into success and failure. Although the difference is small, we were able to distinguish phased communication features extracted from updates and comments according to their prediction power. While some phased campaign communication features were found to have more impact on the outcome of campaigns, others were found to have little correlation with the outcome of the campaigns. In comparison to the existing literature, our phased model results are in general far better than the reported results of 73% in ref. [73], 76.7% in ref. [1], 89.7% in ref. [10], 90.28% in ref. [6] and less than the result of 94.29% reported in ref. [12] of which the update and comment features were not used at all in their optimally weighted Random Forests model. However, the classification results depend on the classifier used since classifiers may have different performances on a particular dataset. It is therefore hard to tell the best classifier given a particular dataset [29]. The intention of this work was to assess the success impact of phased communication features compared with the case involving features that had data accumulated to cover the entire campaign period. Table 4 is the confusion matrix, which further shows the prediction power of our phased model. The table shows the prediction performance of the two models; the phased model and the baseline model. In comparison, the phased model has lower false positive and false negative rates.
than the baseline model, which means that it was making fewer incorrect predictions. The baseline model obtained lower true positive rate of 0.881 and true negative rate of 0.937 compared to our phased model, which had 0.897 true positive rate and 0.939 true negative rate. The results show that the baseline model made more prediction errors compared to our phased model. This accounts for its lower prediction performance than our phased model while the fewer prediction errors of our phased model account for its high prediction performance.

Table 4. Confusion matrix for the two models.

|                | Baseline Model | Phased Model |
|----------------|----------------|--------------|
| Predicted      |                |              |
| -1             | 0.937          | 0.063        |
| 1              | 0.119          | 0.881        |
| Actual         |                |              |
| -1             | 0.937          | 0.063        |
| 1              | 0.119          | 0.881        |

We then computed the average evaluation metric values for prediction of the campaigns’ outcomes for the two models. Our phased model predicts success of campaigns with an accuracy of 89.7%. The result was higher compared to the baseline model by 1.6%. This result shows that our phased model is more accurate than the baseline model in predicting the success of Kickstarter project campaigns. This fact is also indicated by the precision value obtained by our phased model, which shows that of all projects predicted to be successful, 93.6% was the correct prediction. Success prediction performance for all models with regard to different metrics is given in Table 5. The overall results highlight the importance of campaign phases and communication features and thus provides more insights to project creators to focus on the important phases and communication aspects so as to boost the success likelihood of campaigns. Creators of projects can use the identified phases and communication aspects to make necessary efforts to reduce information asymmetry to backers. The results will guide creators of projects to figure out when to communicate and what to emphasize through updates and comments. Moreover, creators can figure out when to encourage conversation through other communication channels like Facebook and Twitter that will trigger backers’ attention on the proposed projects. By focusing on the important phases and communication aspects, creators can manage their time well in planning updates and other project-related endeavors to maximize the funding likelihood. Backers can use the identified phases and communication features to learn about the quality and potentials of the proposed projects to help them make informed funding decisions. Crowdfunding stakeholders can monitor the progress of campaigns and estimate the likely outcome of project campaigns by assessing communication aspects in the identified important phases.

Table 5. Evaluation metrics for the two models.

|                | Successful Campaigns | Failed Campaigns |
|----------------|-----------------------|------------------|
|                | Accuracy   | Recall  | Precision | F-Score | Accuracy   | Recall  | Precision | F-Score |
| Baseline Model | 88.1       | 88.1    | 93.3      | 90.6    | 93.7       | 93.7    | 88.7      | 91.1    |
| Phased Model   | 89.7       | 89.7    | 93.6      | 91.6    | 93.9       | 93.9    | 90.1      | 92.0    |

7. Conclusions

This study explored interactions between backers and creators during crowdfunding campaigns to determine the important aspects of communication and phases for success prediction. To mitigate information asymmetric and predict campaign success, various campaign phases and aspects of communication extracted from updates and comments were assessed. We used PSO, a metaheuristic algorithm, and a KNN classifier for assessing the success contribution power of individual aspects of communication extracted from updates and comments in each campaign phase. Using a Kickstarter
dataset, the study revealed four phases that are the most important with a total of five essential aspects of communication. The identified phases and the corresponding communication features are important for mitigating the information asymmetric between creators and backers and for predicting campaign success. The identified aspects of communication include: The number of updates in phase one; the polarity of comments posted in phase two; the polarity of comments and readability of updates posted in phase three; and the polarity of comments occurring in phase five. The obtained results provide insights to the crowdfunding stakeholders about the important aspects of communication and phases that have high contribution to the success of campaigns. The identified phases and communication features provide an opportunity for creators to do what is necessary in the respective phases so as to boost success likelihood of their project campaigns. The creators may use the identified phases and features to make necessary efforts to accomplish the required communication aspects so as to reduce information asymmetric with the crowds. Moreover, the identified phases and features can help creators to properly manage campaigns by focusing on the required communication features in each identified important phase. Backers can use the phases and communication features to learn about the quality and potentials of the proposed projects. Crowdfunding stakeholders may also estimate the likely outcome of project campaigns by assessing the important phases and communication aspects.

The study conducted in this work has a number of limitations. One of the limitations is the use of one crowdfunding platform, Kickstarter, which uses an “all or nothing rule”. We did not experiment on other well-known platforms like IndieGoGo, which uses different rules that may need different campaign preparations strategies. As such, the limitation hinders the generalization of our findings to such platforms. Future research can focus on assessing the correlation of various communication aspects occurring in various campaign phases with the success likelihood of campaigns on crowdfunding platforms that use different rules. Second, we only considered the time-dependent interaction features that were easy to obtain, and left out other difficult-to-retrieve interaction features like those related to campaigns that happen through social media (e.g., Facebook and Twitter). Research that includes other interaction features may provide more insights on the importance of such interactions and can improve the prediction accuracy. Further research is required to assess the importance of other interaction features in various phases. Third, this study only used KNN and PSO as the machine learning algorithm and metaheuristic algorithm, respectively. The use of other machine learning algorithms (e.g., Random forest, Support Vector Machines, or decision tree) and other metaheuristic algorithms like Ant Colony Algorithm and Firefly Algorithm in assessing the impact of phased communication features will be an interesting future research.

Author Contributions: Conceptualization, M.J.R.; methodology, Y.J. and D.Q.; writing—original draft preparation, M.J.R.; writing—review and editing, S.Q. and Y.J.; project administration, S.Q. and D.Q. All authors have read and agreed to the submission of manuscript for publication considerations.

Funding: This research was funded by the National Social Science Foundation of China, grant number 17BGL083.

Conflicts of Interest: The authors declare no conflict of interest and that the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References
1. Kaur, H.; Gera, J. Effect of Social Media Connectivity on Success of Crowdfunding Campaigns. Proc. Comput. Sci. 2017, 122, 767–774. [CrossRef]
2. Mollick, E.R.M. The dynamics of crowdfunding: An exploratory study. J. Bus. Ventur. 2014, 29, 1–16. [CrossRef]
3. Belleflamme, P.M.; Lambert, T.; Schwienbacher, A. Crowdfunding: Tapping the right crowd. J. Bus. Ventur. 2014, 29, 585–609. [CrossRef]
4. Fernandez-Blanco, A.; Balsera, J.V.; Montequin, V.R.; Moran-Palacios, H. Key Factors for Project Crowdfunding Success: An Empirical Study. Sustainability 2020, 12, 599. [CrossRef]
5. Zvilichovsky, D.; Inbar, Y.; Barzilay, O. Playing Both Sides of the Market: Success and Reciprocity on Crowdfunding Platforms. *SSRN Electron. J.* 2013. [CrossRef]

6. Ryoba, M.J.; Qu, S.; Zhou, Y. Feature subset selection for predicting the success of crowdfunding project campaigns. *Electron. Mark.* 2020, 1–14. [CrossRef]

7. Niemand, T.; Kraus, S.; Angerer, M.; Thies, F.; Turd, A.M. More is not always better—Non-linear effects in crowdfunding. *Int. J. Qual. Innov.* 2019, 5, 6. [CrossRef]

8. Koch, J.-A.; Siering, M. Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms. In Proceedings of the 23rd European Conference on Information Systems (ECIS 2015), Muenster, Germany, 26–29 May 2015; Available online: https://ssrn.com/abstract=2808424, (accessed on 15 June 2020).

9. Xu, A.; Yang, X.; Rao, H.; Fu, W.-T.; Huang, S.-W.; Bailey, B.P. Show me the money! In Proceedings of the 32nd annual ACM conference on Human factors in computing systems—CHI ’14, Association for Computing Machinery (ACM), Toronto, ON, Canada, 26 April–1 May 2014; pp. 591–600. [CrossRef]

10. Chen, S.-Y.; Chen, C.-N.; Chen, Y.-R.; Yang, C.-W.; Lin, W.-C.; Wei, C.-P. Will Your Project Get the Green Light? Predicting the Success of Crowdfunding Campaigns. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Singapore, 6–9 July 2015; p. 79.

11. Lai, C.-Y.; Lo, P.-C.; Hwang, S.-Y. Incorporating Comment Text into Success Prediction of Crowdfunding Campaigns. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Langkawi, Malaysia, 16–20 July 2017; p. 156.

12. Ahmad, F.S.; Tyagi, D.; Kaur, S. Predicting crowdfunding success with optimally weighted random forests. In Proceedings of the 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, UAE, 18–20 December 2017; pp. 770–775.

13. Miglo, A.; Miglo, V. Market imperfections and crowdfunding. *Small Bus. Econ.* 2018, 53, 51–79. [CrossRef]

14. Liang, X.; Hu, X.; Jiang, J. Research on the Effects of Information Description on Crowdfunding Success within a Sustainable Economy—The Perspective of Information Communication. *Sustainability* 2020, 12, 650. [CrossRef]

15. Block, J.H.; Hornuf, L.; Moritz, A. Which Updates During an Equity Crowdfunding Campaign Increase Crowd Participation? *SSRN Electron. J.* 2016, 50, 3–27. [CrossRef]

16. Friedkin, N.E. *A Structural Theory of Social Influence*; Cambridge University Press: Cambridge, UK, 1998; Volume 13.

17. Sheng, J. Being Active in Online Communications: Firm Responsiveness and Customer Engagement Behaviour. *J. Interact. Mark.* 2019, 46, 40–51. [CrossRef]

18. Talbi, E.-G. A Taxonomy of Hybrid Metaheuristics. *J. Heuristics* 2002, 8, 541–564. [CrossRef]

19. Crawford, B.; Soto, R.; Astorga, G.; Garcia, J.; Castro, C.; Paredes, F.; Garcia, J. A Putting Continuous Metaheuristics to Work in Binary Search Spaces. *Complexity* 2017, 2017, 1–19. [CrossRef]

20. Osman, I.H.; Laporte, G. Metaheuristics: A bibliography. *Ann. Oper. Res.* 1996, 63, 511–623. [CrossRef]

21. Poli, R.; Kennedy, J.; Blackwell, T. Particle swarm optimization. *Swarm Intell.* 2007, 1, 33–57. [CrossRef]

22. Kennedy, J.; Eberhart, R.C. A discrete binary version of the particle swarm algorithm. In Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, Orlando, FL, USA, 12–15 October 1997; Volume 5, pp. 4104–4108.

23. Yadav, S.; Ekbal, A.; Saha, S. Information theoretic-PSO-based feature selection: An application in biomedical entity extraction. *Knowl. Inf. Syst.* 2018, 60, 1453–1478. [CrossRef]

24. Chuang, L.-Y.; Chang, H.-W.; Tu, C.-J.; Yang, C.-H. Improved binary PSO for feature selection using gene expression data. *Comput. Biol. Chem.* 2008, 32, 29–38. [CrossRef]

25. Cervante, L.; Xue, B.; Zhang, M.; Shang, L. Binary particle swarm optimisation for feature selection: A filter based approach. In Proceedings of the 2012 IEEE Congress on Evolutionary Computation, Brisbane, QLD, Australia, 10–15 June 2012; pp. 1–8.

26. Wang, X.; Yang, J.; Teng, X.; Xia, W.; Jensen, R. Feature selection based on rough sets and particle swarm optimization. *Pattern Recognit. Lett.* 2007, 28, 459–471. [CrossRef]

27. Zhang, Y.; Gong, D.-W.; Hu, Y.; Zhang, W. Feature selection algorithm based on bare bones particle swarm optimization. *Neurocomputing* 2015, 148, 150–157. [CrossRef]
28. Aljarah, I.; Al-Zoubi, A.M.; Faris, H.; Hassonah, M.A.; Mirjalili, S.; Saadeh, H. Simultaneous Feature Selection and Support Vector Machine Optimization Using the Grasshopper Optimization Algorithm. *Cogn. Comput.* 2018, 10, 478–495. [CrossRef]
29. Harrington, P. *Machine Learning in Action*; Manning Publ. Co.: New York, NY, USA, 2012.
30. Petruzzielli, A.M.; Natalicio, A.; Umberto, P.; Roma, P. Understanding the crowdfunding phenomenon and its implications for sustainability. *Technol. Forecast. Soc. Chang.* 2019, 141, 138–148. [CrossRef]
31. Vismara, S. Sustainability in equity crowdfunding. *Technol. Forecast. Soc. Chang.* 2019, 141, 98–106. [CrossRef]
32. Cumming, D.; Leboeuf, G.; Schwienbacher, A. Crowdfunding cleantech. *Energy Econ.* 2017, 65, 292–303. [CrossRef]
33. Bento, N.; Gianfrate, G.; Thoni, M.H. Crowdfunding for sustainability ventures. *J. Clean. Prod.* 2019, 237, 117751. [CrossRef]
34. Harrington, P. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003, 3, 1157–1182. [CrossRef]
35. Bento, N.; Gianfrate, G.; Groppo, S. Do crowdfunding returns reward risk? Evidences from clean-tech projects. *Technol. Forecast. Soc. Chang.* 2019, 141, 107–116. [CrossRef]
36. Hörisch, J. Think Big or Small is Beautiful. An empirical analysis of characteristics and determinants of success of sustainable crowdfunding projects. *Int. J. Entrep. Ventur.* 2018, 10, 1. [CrossRef]
37. Stiglitz, J.E. The Contributions of the Economics of Information to Twentieth Century Economics. *Q. J. Econ.* 2000, 115, 1441–1478. [CrossRef]
38. Courtney, C.; Dutta, S.; Li, Y. Resolving Information Asymmetry: Signaling, Endorsement, and Crowdfunding Success. *Entrep. Theory Pract.* 2016, 41, 265–290. [CrossRef]
39. Akerlof, G.A. The market for ‘lemons’: Quality uncertainty and the market mechanism. In *Uncertainty in Economics*; Elsevier: Amsterdam, The Netherlands, 1978; pp. 235–251.
40. Ahlers, G.K.; Cumming, D.J.; Guenther, C.; Schweizer, D. Signaling in Equity Crowdfunding. SSRN Electron. J. 2012, 39, 955–980. [CrossRef]
41. Gerber, E.M.; Hui, J.S.; Kuo, P.-Y. Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms. In Proceedings of the International Workshop on Design, Influence, and Social Technologies: Techniques, Impacts and Ethics, Evanston, IL, USA, 11–15 February 2012; Volume 2, p. 10.
42. Spense, M. Job market signaling. *Q. J. Econ.* 1973, 87, 355–374. [CrossRef]
43. Kang, H.; Kim, H.U. Who Can Survive in an ICT-Enabled Crowdfunding Platform? *Sustainability* 2020, 12, 504. [CrossRef]
44. Adger, W.N. Social and ecological resilience: Are they related? *Prog. Hum. Geogr.* 2000, 24, 347–364. [CrossRef]
45. Davies, W.E.; Giovannetti, E.G. Signalling experience & reciprocity to temper asymmetric information in crowdfunding evidence from 10,000 projects. *Technol. Forecast. Soc. Chang.* 2018, 133, 118–131. [CrossRef]
46. Wang, N.; Li, Q.; Liang, H.; Ye, T.; Ge, S. Understanding the importance of interaction between creators and backers in crowdfunding success. *Electron. Commer. Res. Appl.* 2018, 27, 106–117. [CrossRef]
47. Efrat, K.; Gilboa, S. Relationship approach to crowdfunding: How creators and supporters interaction enhances projects’ success. *Electron. Mark.* 2019, 1–13. [CrossRef]
48. Rao, R.S.; Chandy, R.K.; Prabhu, J.C. The Fruits of Legitimacy: Why Some New Ventures Gain more from Innovation than Others. *J. Mark.* 2008, 72, 58–75. [CrossRef]
49. Wessel, M.; Thies, F.; Benlian, A. Opening the Floodgates: The Implications of Increasing Platform Openness in Crowdfunding. *J. Inf. Technol.* 2017, 32, 344–360. [CrossRef]
50. Lee, S.; Lee, K.; Kim, H.C. Content-based Success Prediction of Crowdfunding Campaigns. In Proceedings of the Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing—CSCW ’18, Association for Computing Machinery (ACM), Jersey City, NJ, USA, 30 October 2018; pp. 193–196.
51. Rao, H.; Xu, A.; Yang, X.; Fu, W.-T. Emerging Dynamics in Crowdfunding Campaigns. In *Proceedings of the Haptics: Science, Technology, Applications*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2014; Volume 8393, pp. 333–340.
52. Talbi, E.-G. *Metaheuristics: From Design to Implementation*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
53. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003, 3, 1157–1182.
54. Mtw; Liu, H.; Motoda, H. Feature Extraction Construction and Selection: A Data Mining Perspective. *J. Am. Stat. Assoc.* 1999, 94, 1390. [CrossRef]

55. Etter, V.; Grossglauser, M.; Thiran, P. Launch hard or go home! In Proceedings of the first ACM conference on Wireless network security—WiSec ’08, Association for Computing Machinery (ACM), Boston, MA, USA, 7 October 2013; pp. 177–182.

56. Elenechev, A. Vasilev, and others, Forecasting the Success Rate of Reward Based Crowdfunding Projects. *ZBW 2017*, 17, 2–34.

57. Wang, W.; Zhu, K.; Wang, H.; Wu, Y. The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics. *IET Softw.* 2017, 11, 229–238. [CrossRef]

58. Xia, Y.; Cambria, E.; Hussain, A.; Zhao, H. Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features. *Cogn. Comput.* 2014, 7, 369–380. [CrossRef]

59. Mafarja, M.; Mirjalili, S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. *Neurocomputing* 2017, 260, 302–312. [CrossRef]

60. Kennedy, J.; Eberhart, R.C.; Shi, Y. *Swarm Intelligence*; Elsevier: Amsterdam, The Netherlands, 2001.

61. Chuang, L.-Y.; Tsai, S.-W.; Yang, C.-H. Improved binary particle swarm optimization using catfish effect for feature selection. *Expert Syst. Appl.* 2011, 38, 12699–12707. [CrossRef]

62. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* 2014, 40, 16–28. [CrossRef]

63. Domeniconi, C.; Peng, J.; Gunopulos, D. Locally adaptive metric nearest-neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002, 24, 1281–1285. [CrossRef]

64. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, SciPy, Austin, TX, USA, 28–30 June 2010; Volume 445, pp. 56–61.

65. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O’Reilly Media Inc.: Champaign, IL, USA, 2009.

66. Gong, M.; Yan, J.; Shen, B.; Ma, L.; Cai, Q. Influence maximization in social networks based on discrete particle swarm optimization. *Inf. Sci.* 2016, 367, 600–614. [CrossRef]

67. Zhou, M.; Lu, B.; Fan, W.; Wang, G. A. Project description and crowdfunding success: An exploratory study. *Inf. Syst. Front.* 2016, 20, 259–274. [CrossRef]