Hybrid Fraud Detection Model: Detecting Fraudulent Information in the Healthcare Crowdfunding

Jaewon Choi[1] , Jaehyoun Kim[2], and Ho Lee[3*]
[1] Department of Business Administration, Soonshunhyang University
22 Soonchunhyang-ro, Shinchang-Myeon, Asan, Chungnam 335-745 – South Korea
[e-mail: jaewonchoi@sch.ac.kr]
[2] Department of Computer Education, Sungkyunkwan University
25-2, Sungkyunkwan-ro Jongno-gu – South Korea
[e-mail: jaekim@skku.edu]
[3] Department of Future Technology, Korea University of Technology and Education
1600, Chungjeol-ro, Cheonan-si, Chungnam 31253 – South Korea
[e-mail: leeho32@koreatech.ac.kr]
[*]Corresponding author: Ho Lee

Received October 23, 2021; revised December 15, 2021; accepted February 5, 2022; published March 31, 2022

Abstract

In the crowdfunding market, various crowdfunding platforms can offer founders the possibilities to collect funding and launch someone's next campaign, project or events. Especially, healthcare crowdfunding is a field that is growing rapidly on health-related problems based on online platforms. One of the largest platforms, GoFundMe, has raised US$ 5 billion since 2010. Unfortunately, while providing crucial help to care for many people, it is also increasing risk of fraud. Using the largest platform of crowdfunding market, GoFundMe, we conduct an exhaustive search of detection on fraud from October 2016 to September 2019. Data sets are based on 6 main types of medical focused crowdfunding campaigns or events, such as cancer, in vitro fertilization (IVF), leukemia, health insurance, lymphoma and, surgery type. This study evaluated a detect of fraud process to identify fraud from non-fraud healthcare crowdfunding campaigns using various machine learning technics.

Keywords: Crowdfunding, Fraud Detection, Collaborative Filtering, Social SVD, LDA

This research was supported by the Soonchunhyang University Research Fund.
This paper was supported by Education and Research promotion program of KOREATECH in 2022.

http://doi.org/10.3837/tiis.2022.03.014 **ISSN : 1976-7277**

1. Introduction

Crowdfunding is a way for people, companies, and charities to raise money from crowd. It works through individuals or organizations investing (or donating) in crowdfunding campaigns or projects in return for a potential benefits or reward. Crowdfunding platforms provide project or campaign founders with opportunities to raise funds for the realization of projects or campaigns. Over the past few years, crowdfunding platforms have provided campaigns as online services to provide opportunities to raise funds or money for projects and ideas. Crowdfunding platforms are becoming increasingly important in financing projects of various sizes [1]. Kickstarter, a leading platform, attracted more than 250,000 projects with a total of $1.9 billion in funds and had a success rate of about 35% by 2015 [1]. In addition, the trading value of the crowdfunding sector amounts to USD 692.36 million in 2019. The transaction value is expected to grow at 14.7% per year, and by 2023, it will be a total of $198.6 million (Statista.com).

Despite the popularity of crowdfunding platforms, there is an increasing risk of cheating in campaigns and projects prevalent in crowdfunding platforms and some open online services. As a result, the crowdfunding platform introduces an integrity plan to carry out projects posted or uploaded on the crowdfunding platform and intervene in fraudulent processes. Fraudulent processes or actions reduce opportunities for honest campaigns or project founders and lead to a state of non-attention from potential campaigns or project funders working on the platform. In addition, if a fraudulent campaign or project is finally fully supported, it is questionable (1) whether the main goals of the project can be achieved, and (2) whether rewards or donations will finally be delivered to the campaign or project cost applicants. Therefore, it is important to detect cheating on crowdfunding platforms.

Previous research of fraud detection in crowdfunding market have focused on the identifying static fraud website content [2] as well as fraudulent computer-mediated communication, which occurs dynamically between widely different agents [3]. The new platform environment in the crowdfunding market combines research streams and provides opportunities to detect data and text-based fraud. Campaigns or project founders can post information about campaigns or projects such as images or videos through crowdfunding platforms. In addition, they can communicate or interact with potential funders through other channels. Previous studies have relied on content-based and verbal cues to identify fraud in text-type unstructured data. [4]. On one hand, linguistic cues are derived from texts and encompass abstract aspects such as expressed emotionally or characteristics are like the diversity of the used vocabulary[3]. Meanwhile, in the previous research, content-based cues are directly related to the texts surrounding the single used words expressed by a text mining method, such as frequency of a bag-of-words.

However, existing studies have mainly focused on questions related to factors that ignore questions about how to induce successful financing of crowdfunding campaigns or projects and identify suspicious projects [5]. Previous studies have mainly focused on factors that lead to successful campaigns or, conversely, only considered factors that can distinguish fraudulent campaigns. However, in this study, both successful and deceptive campaign elements were considered simultaneously. In addition, this study focuses on expanding previous understanding by labeling areas of trick detection on crowdfunding platforms such as healthcare-based crowdfunding platforms. Therefore, this study collects and analyzes word data sets from campaigns posted on the GoFundMe platform. We propose and evaluate various methods that can help with fraud detection mechanisms. In addition to contributing to the field of crowdfunding research, this study also contributes to the general field of fraud detection by

combining the field of healthcare crowdfunding with various types of data analytics of detecting fraud on healthcare crowdfunding platforms. Thus, the main research questions derived from the above discussion are summarized as follows. First, how is the automation method determined to identify fraudulent healthcare crowdfunding campaigns in healthcare crowdfunding campaigns? Second, can the text information of crowdfunding provided by Founder be used to detect fraudulent campaigns? Third, are Latent Dirichlet allocation (LDA) and Collaborative Filtering (CF) effective in detecting fraud in crowdfunding campaigns?

2. Literature Review

2.1 Crowdfunding Market and Healthcare Field

Healthcare crowdfunding is a rapidly becoming common campaign category in many popular and leading crowdfunding platforms. In addition, it has become a popular choice worldwide for people with unaffordable health needs and is being made on donation-based platforms [6]. The crowdfunding model provides a great feature for short-term economic problems facing individuals from medical-related treatment costs and other related needs. One of the biggest examples of platforms that enable such donation-based crowdfunding is GoFundMe. For example, healthcare campaigns, especially on the GoFundMe platform, increased sevenfold between 2011 and 2014. The healthcare campaigns raised more than $930 million [7]. In a similar way, the GiveForward platform, a healthcare-oriented crowdfunding platform merged with YouCarring, has recently raised $200 million.

In traditional healthcare fundraising, charities have relied on third parties and have cautiously created their reputation over time to promote their credibility to potential donors [8]. As healthcare crowdfunding campaigns move to online platforms such as GoFundMe and others, potential donors (backers) can no longer rely on themselves. Healthcare-related crowdfunding platforms have the full potential to reducing medical related financial issues and about an actual social impact by marking healthcare related insurance gaps. The overall crowdfunding movement has generally been considered successful by democratizing entrepreneurs' access to capital [9].

Table 1. Literature with a Perspective Crowdfunding

Field	Content	References
Information System	- The relationship between healthcare crowdfunding and national health systems - Medical- related crowdfunding campaigns have a significant, negative impact on the personal bankruptcy - Examined the crowdfunding from an HCI perspective- especially from a motivational perspective and looking at both the creator and the funder	[7, 10-12]
Management	- An alternative paradigm of societal problem solving - Analysis of dynamic communication during the crowdfunding period is valuable for identifying fraudulent behavior	[1, 13, 14]

Healthcare	- The case of Charlie Gard serves as a useful example to reveal the hidden impacts of crowdfunding on healthcare and its delivery, as the extraordinary success of the campaign draws attention to ethical concerns - Successful crowdfunding approach at a non-profit organization is getting funding for rare disease research - Crowdfunding for cancer-related care is likely to benefit	[15-17]

Most campaign creators collect donations from the public through online platforms (Facebook, Twitter, e-mail, etc.) for sharing and promotion. Funding for the GoFundMe platform is divided into seven areas: breast cancer, cancer, IVF (In Vitro Fertilization), leukemia, lymphoma, health insurance and surgery. Raising research funds through crowdfunding is more difficult than developing creative projects. Scientific knowledge is necessary to understand healthcare campaigns that are difficult to communicate with heterogeneous people. As a result, there is a special crowdfunding platform that can be posted only for healthcare-scale campaigns. **Table 1** described the research stream is based on both the management and healthcare sectors.

Crowdfunding allows all types of entrepreneurs, including art, culture, society, and for-profit, to raise funds from the crowd [12]. The crowdfunding market is usually greatly encouraged by the success of the crowdfunding paradigm through Internet platforms [18]. There are four types of crowdfunding[18, 19], generally including donation-based crowdfunding, reward-based crowdfunding, lending crowdfunding and equity-based crowdfunding. First, the donation-based crowdfunding type promotes private donations to public goods, and some donations are based on altruism [19]. Non-profit organizations are using crowdfunding to raise donations. Individuals use personal crowdfunding on platforms such as GoFundMe, YouCaring, or other platforms to raise funds to solve personal problems for themselves, friends, and family. This approach is often effectively used to deal with personal tragedies, whether a child has cancer, parents have other economic crises or other related problems. Funders (people or crowds) often react generously to such campaigns, which seem to prefer to give money directly to someone, the founder of campaigns in need, rather than donating to organizations that provide support to the same family. Second, in a reward-based crowdfunding model, donations will earn "community benefits" [20] in return for financial participation. Kickstarter, a famous reward-based platform, is used by entrepreneurs and inventors to finance the development and production of campaigns or projects. In a same way, IndieGogo platform, which describes itself as a crowdfunding site, is popular for many purposes, such as Kickstarter. The former allows a wide range of campaigns on the site and does not critically require the use of whole or no campaigns. The all or nothing approach required by Kickstarter means that campaigns or projects that have not reached their goal are not paid at all for their pledges and the fundraiser receive nothing. Third, lending-based crowdfunding type is considered P2P loans to individuals, groups, or SMEs, and are expected to be repaid in interest or interest-free after a certain period of time [21]. Individuals or companies running campaigns are essentially looking for borrowing money from the crowd. In other way, debt-based crowdfunding type campaigns are particularly popular with entrepreneurs who do not want to give up their stake in startups immediately and do not have access to more traditional types of loans or debt facilities. Finally, in the equity-based crowdfunding type, investors become stakeholders and receive dividends according to the company's investment performance. The term crowdfunding is written into the title of the governing rules from the Securities and Exchange Commission (SEC) in USA. Regulation crowdfunding establishes the code for both issuers and the intermediaries (platforms and

broker-dealers). Also, FINRA, the entity authorized by the SEC to regulate crowdfunding, lists the platforms engaged in the business on its website.

On a crowdfunding platform, a campaign or project founder can raise funds from potential funders (i.e., platform users) to realize a campaign or project. Through the crowdfunding platform, the general public can make small investments in ventures promoted by entrepreneurs [22]. During the crowdfunding process, founders and fund providers generally communicate with a rich set of information related to campaign or project ideas that provide funds by switching users on the platform [23]. Previous studies have shown that campaign or project characteristics, such as the fund's target amount, funding period or campaign period, and the founder's number of Facebook friends, affect the successful funding of the project [12]. In addition, previous studies have in some context provided the influence of profit-orientation [24] or campaign donation motivation on crowdfunding platforms [25]. Most of the announced campaigns or projects are fully funded when the target amount is achieved during the funding period of the campaign or project or at the end of the funding period. Nevertheless, after the funding period of a campaign or project, some projects are suspended, deactivated, or deactivated due to founder fraud.

As the number of crowdfunding platforms increases, the risk associated with cheating on campaigns or projects prevalent in crowdfunding platforms and some open online services is increasing. In particular, the risk of fraud related to the crowdfunding market is also contrary to the back-responsibility of the crowdfunding type. Equity-based and lending-based crowdfunding has higher levels of responsibility than donation-based and reward-based types. The existence of fraud campaigns or projects is disadvantageous not only to the platform itself, but also to the owner (stakeholder) of the crowdfunding platform. If a fraudulent campaign or project is not detected and has been successfully funded, the founder may not be able to deliver the proposed project results, which may result in a loss of investment by the campaign or project funding provider. In addition, if fraudulent campaigns or projects gain popularity in the media, other normal projects will not receive attention or face problems that are difficult to fund without fraud. Finally, the occurrence of fraud on the platform reduces the reliability and reliability of the crowdfunding platform itself, which can work with sub-platform users and pose significant risks to platform operators.

2.2 Fraud Detection

Fraud occurs by many different types of forms. These categories are divided into economic or financial activities that are vulnerable to fraud or other environmental, technological, or fraudulent approaches. The detection model to the cycle of fraud and the four essential activities as follow: (1) Fraud detection mean applying detection models on new or new released observations related to assigning a fraud risk to every observation. (2) Fraud investigation: A human expert is often required to investigate suspicious, flagged cases given the involved subtlety and complexity. (3) Fraud confirmation: Determining true fraud new or label. (4) Fraud prevention: Preventing process of determining the fraud process is full possibly to involving field research. Traditionally, fraud and deception have been extensively studied in the field of face-to-face communication and computer media communication among social scientists. On top of that, fraudulent agents disclose fraud information through credit card, banking, and insurance fraud[26, 27], fraudulent news[28], fraudulent corporate communication[29]; and false sites[30].

Some of the previous studies focused on detecting fraud in the crowdfunding market and studies related to fraud detection in crowdfunding are still lacking. The detection of deception and fraud is challenging task. Previous studies have examined content-based cues (e.g., words

in text) and linguistic cues (e.g., positive or negative effects, lexical complexity, diversity and specificity, quantity terms, uncertainty and informality). Content-based cues focus on real content, such as "What is delivered?" and linguistic cues focus on the question "How is deception delivered in natural language?"[4]. Content-based cue approach is mainly based on text and bag-of-words in that text. However, classification algorithms, content-based and linguistic based cues are for reward-based crowdfunding[31].

Both types of cues are extracted from static and dynamic communication as shown in **Table 2**. Various studies on financial industries such as banks, insurance, credit card transactions, and many other industries and fields have been analyzed in several ways. However, there is still a lack of detection of crowdfunding fraud, such as the main process of medical crowdfunding campaigns. Through this study, we determined to close these research gaps and add value to specific areas of fraud crowdfunding campaigns or projects as well as to general fraud detection areas considering previous theoretical studies as shown in **Table 2**.

Table 2. A Comparison of Theories related to Fraud Detection

Theory/ Model	Emphasis of Communication Perspective		Verbal Cues	
	Static	Dynamic	Linguistic	Content-based
Channel Expansion Theory (CET)	+	+	+	
Competence Model of Fraud Detection (CM)	+			+
Criteria-Based Content Analysis (CBCA)	+		+	
Four- Factor Theory (FFT)	+		+	
Information Manipulation Theory (IMT)	+		+	
Interaction Adaption Theory (IAT)		+	+	
Interpersonal Deception Theory (IDT)		+	+	
Leakage Theory (LT)	+		+	
Reality Monitoring (RM)	+		+	
Scientific Content Analysis (SCAN)	+		+	

Channel Expansion Theory (CET) is an extension of media richness theory[32, 33], which is a communication media identification theory that integrates the empirical elements that predict and explain users' perceptions of new communications. CET suggests that users gain more experience and knowledge by using channels and try to explain how various forms of electronic communication affect their perception. In order to detect fraudulent events, such as financial and accounting audits, the Competence Model of Fraud Detection (CM) was tested and exemplified in the context [34]. The model supposes that the deceiver uses deception tactics such as masking, dazzling, decoying, repackaging, mimicking, and double play [35]. Also, Criteria-Based Content Analysis (CBCA) is a key component of syntax validation (SVA) technology and is widely used worldwide to detect verbal account spoofing in various contexts [36]. CBCA is a forensic tool for determining authenticity.

Four-Factor Theory (FFT) identifies four basic psychological factors (arousal, attempted control, affective response and cognitive factors) in the event of an anomaly or deception [37]. Deceivers feel excited and anxious (e.g., duplication of words, hesitation, etc.) when lying, such as when the human nervous system naturally reacts. Similarly, Leakage Theory (LT) had been proposed as a psychiatric theoretical model[38]. According to Leakage theory, emotional reactions must occur through deception, and such changes must affect behavior. In previous

studies, there were five types of channels for leakage theory: body language, facial expression, language style, voice, and language content.

Information Manipulation Theory (IMT) is a method of observing interpersonal communication and interpreting deception from a communication perspective [39]. IMT suggests that fraudulent information has a fraudulent function because it violates the principles of communication. On the other hand, one of the four conversations was deliberately interrupted to deceive. First, the quality of information (information is truthful and correct) is at best questionable, that is, the amount of information is sufficient and not missing. The information is also presented in a way that is relevant to the topic of conversation at hand and understandable at the end. A fraudulent activity or project provider that uses these methods when sharing an activity or project description or communicating with a potential funding provider in the context of crowdfunding.

Table 3. The Categorization of Verbal Cues Research

Category	Feature		Theory
Linguistic Cues	Affect	- Affect ratio - Pos. affect - Neg. affect - Pleasure	FFT, IDT, LT
	Complexity	- Avg. sentence Length - Avg. word Length - Noof. Clauses - Pausality	IDT, IMT
	Diversity Expressivity	- Lexical diversity emotiveness	IDT IDT
	Non-immediacy	- Group reference - Individual references - Self references	IDT
	Quantity	- Modifier quantity - Sentence quantity - Verb quantity - Word quantity	IDT, IMT
	Specificity	- Perceptual information and sensory ratio	RM, IDT
	Uncertainty	- Modal verb ratio - Uncertainty words	IDT, IMT
	Informality	- Typographical error ratio	IDT
Content-Based Cues	-	- A-bag-of-words	CM

Interaction Adaption Theory (IAT) describes how individuals change their behavior according to the behavior of others in a conversation [40]. There are three main factors contributing to the IAT: requirements, expectation and desires. The key is that when all messages are viewed summarily, fraudulent information is periodically embedded in real information, or interaction adaptation theory hypothesizes that all message combinations are better than parsing individual messages to capture basic dynamics. Like the IAT, The Interpersonal Deception Theory (IDT) interprets spoofing in personal communication as a message exchange based on a communication perspective and focuses on intentional and strategic communication between the sender and receiver [41]. Real-life monitoring has been studied, and more generally, source monitoring influences false factors. Reality Monitoring (RM) [34] and Scientific Content Analysis (SCAN) [42] are related other techniques that also rely on truth evaluation based on

suspects' written statements by linguistic cues. Reality monitoring theory suggests that real-life surveillance can be based on what is memorized and how independent the general dimension is. SCAN theory is currently used as a global and linguistic integrity evaluation model and is used to analyze statements and determine whether to be the author of the communication as shown in **Table 3**.

As a result, various techniques of fraud detection can help identify fraud campaigns or projects using published information. Campaign creator or project founders have different opportunities to announce campaigns or projects and to communicate with their (potential) funders. Platform users can post comments or choose whether to ask questions and review information on campaigns or projects. The main source of campaign or project-related information is the project page. The page displays basic information, such as the target amount, and funding period and key information. There are also several possibilities for dynamic communication, including other users of the platform, after the funding period for the campaign or project begins. First, platform users can leave comments or ask questions about the campaign project, and campaign authors or project founders can post responses to their exploration on the comment page. And another source of dynamic communication is page updates. Here, the campaign creator or project founder can change updates related to the status of the campaign or project. Previous studies in the field of crowdfunding do not focus on fraud, especially fraud occurring on healthcare-related platforms. However, due to the nature of the crowdfunding platform, fraud represents an important problem to consider on the Internet, characterized by general features such as low entry barriers, spatial and temporal separation, and high anonymity to market participants. Against this background, this study has an incentive to provide an appropriate fraud detection mechanism through text and data information of healthcare campaigns to solve the problem of the crowdfunding market.

3. Research Framework

In this section, we proposed Fraud Detection Model (FDM) for detecting fraud against crowdfunding campaigns, especially healthcare-related crowdfunding activities such as **Fig. 1**.

First, preprocessing processes such as raw data are called source data. In this study, raw data sets are randomly collected from the medical category of the GoFundMe platform. All datasets include healthcare crowdfunding campaign's article, story, date, goal amount, raised amount and number of donors raised in the medical crowdfunding campaign. Next step is process missing values. The purpose of missing values is very important for understanding the order of successfully managed data. After handling missing value, normalization and data separation will be next step. After processing the missing values, normalization and data separation proceed to the next step. The purpose of data normalization is to reduce and even eliminate data duplication when a predictor (or set of predictors) fully predicts the outcome variable.

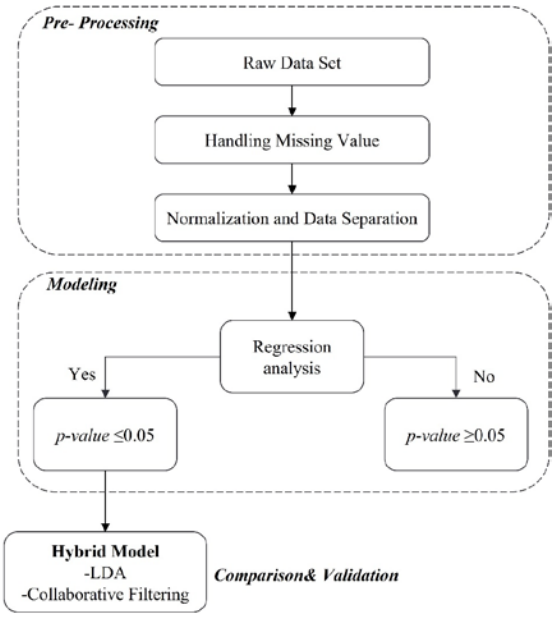


Fig. 1. Framework for Fraud Detection Model (FDM)

Next section is modeling process. The purpose of regression is a series of statistical processes to estimate the relationship between variables. The reason why regression was chosen to model the process is to predict the results based on historical data collected from the campaign's data on the GoFundMe platform. It helps to create a model that predicts results based on a regression model. Next process is determining the p-value. The p-value is the level of marginal significance representing the probability of the occurrence of a given event. As we known as, in the majority analysis, an alpha of 0.05 is used as the cut-off for significance. In this study, the GoFundMe platform's medical campaign target amount, increased amount, and donor information are used to perform regression analysis and determine p-values greater than 0.05 or less than 0.05. If the p-value is less than 0.05 (typically ≤0.05), go to the next step.

The final step is to compare and validate the process through hybrid mode. This hybrid model mixes LDA (Latent Dirichlet Ascension) and collaborative filtering for fraud medical crowdfunding campaigns. In this study, textual information about medical activities, such as GoFundMe articles or stories, is used to compare and validate the process.

3.1 Latent Dirichlet Allocation (LDA)

Natural language Processing (NLP) was used in to detect fraud in the telecommunications industry[43] such as to build user profile signatures. Among NLP method, Latent Dirichlet Allocation (LDA) is innovative technic to identify psychological clustering from huge text information. It assumes that any significant unexplainable deviations from the normal activity of an individual user is strongly correlated with fraudulent activity.

LDA is processes document as a bag-of-words. The bag-of-words model is one of the most widely used models in object categorization and the main idea is to define each extracted key point as one of the visual words and then represent each image as a histogram of the visual word [44]. Bag-of-words model is used to train the frequency of occurrence of each word in a file categorization method. Bag-of-words model has been of interest for numerous computer vision tasks ranging from discover [45] and object classification [46] to object retrieval [47] and texture classification[44]. Several methods for detecting telecommunication fraud are also

proposed. Also, those methods is used to detecting automobile insurance fraud [48], fraud in financial report[49], fraud in health insurance [50] and other industry field. For example, former researchers introduced LDA to fake review detection problems [51-53]. However, their methods apply LDA directly upon review text[52]. It is effective to apply LDA by comparing subtle differences between fraudulent and true reviews [54]. Crowd campaigns that achieve significant exposure can significantly increase donations compared to other activities targeting individuals and large-scale social networks or increased links between news reports and crowdfunding campaigns. Therefore, the LDA method is used to compare and validate the fraud detection model, one of the purposes of this study.

3.2 Collaborative Filtering (CF)

Collaborative filtering aims to provide personalized items based on past interactions such as clicks and ratings [55]. Collaborative filtering algorithms have been one of the best methods for recommender systems and are one of the key tools to provide relevant content and drive successful electronic commerce[56]. This algorithm is a system that collects the tendencies or tastes of numerous users and automatically predicts users' preference. Recommender systems are beneficial to both service providers and users. It reduces transaction costs of finding and selecting items in an online shopping environment [57, 58]. Previous studies have proven that CF improves the decision-making process and quality. Previous studies suggested the recommendation system was defined as a decision-making strategy for users in complex information environments, which helped users retrieve records of knowledge related to the user's interest and preferences [59, 60].

Recently, various approaches have been developed to establish a recommended system that can utilize content-based filtering [61], user-to-user collaborative filtering [62] and hybrid filtering[63]. In the content-based approach, the utilities assigned by the same user to other items are similar to systems like recommending items by containing articles and some news or messages, textual information such as websites or platforms[64]. Other research suggested the systems which usually consider content-based approach or filtering to help their users to find right information on the Internet[65]. The system uses a user interface that helps users navigate the Internet. This interface may track a user's search pattern to predict a page of interest [66]. On the other hand, user-to-user collaborative filtering distinguishes users with similar interests or tastes from other users and recommends items to active users. For example, GroupLens has adopted a collaborative method that supports users to find articles in large news databases [67]. Despite the advantages of these two CF filtering techniques, several limitations have been identified in content-based filtering, such as limited content analysis, overspecialization and sparsity of data[67]. Also, user-to-user CF exhibits cold start, scarcity and scalability problems. These problems usually degrade the quality of recommendations. To reduce some of the identified issues, the hybrid filtering combines two or more approaches to increase the accuracy and performance of recommender system in other ways [68]. Hybrid approaches integrate both content-based and user-to-user CFs to achieve further improved recommendations. The predictive recommender systems can be improved when mixing multiple predictors by substantially and that would be rather than using one or single approach or method [69].

In the subclassification of the CF algorithm, there are memory-based algorithms. Memory-based approach is essentially use heuristics to rating predictions based on the entire collection of items, that are previously rated by active users [64]. That is, the unknown rating of an item-user combination can be estimated as an aggregate of ratings of the most similar users for the

same item. Traditionally, model-based algorithms have been used to alleviate the scalability problems associated with memory-based recommender systems. Memory-based algorithms are divided into User-to-user and Item-based in the neighborhood model, and Model-based algorithms are divided into Matrix Factorization (hereinafter referred to as MF), RBM, and Bayesian. The method of inferring the empty space of the rating from a memory-based algorithm has the same problem as the MF but has a different solution. Memory-based algorithms fill empty spaces using similar users or similarity coefficients of items, or solve them by making them a regression problem. The main drawback of memory-based technique is the requirement of loading a large amount of in-line memory [70]. Model-based approach intends to solve such problems. Among model-based algorithms, the most widely used MF is a method of decomposing or reducing vectors representing user or product information into algorithms such as PCA(Principal Component Analysis) or SVD(Singular Value Decomposition). Therefore, this study judged that it was necessary to use both memory-based and model-based approaches for the extraction of fraudulent content.

4. Research Methodology

4.1 Data Collection and Procedure

The data used in the analysis were collected from GoFundMe.com, an online crowdfunding platform. On GoFundMe platform, campaigns are related to a variety of topics can be published by campaign creator and funded by the platform's users. The collected data were collected from 10,012 campaigns that occurred during the period from October 2016 to September 2019. As shown in **Table 4** and **Table 5**, there were many campaigns related to Cancer and IVF, and various currencies were used in addition to US Dollar.

Table 4. Characteristics of collected data

Num.	Category	2016.Oct.01-2017.Sep. 30	2017.Oct.01-2018. Sep.31	2018.Oct.01-2019. Sep.30	Total (#)
1	Cancer	637	369	690	1696
2	IVF	431	410	615	1456
3	Leukemia	383	274	816	1473
4	Health Insurance	539	352	731	1622
5	Lymphoma	478	442	657	1577
6	Surgery	965	531	692	2188
Total		3433	2378	4201	10012

Table 5. Campaigns by Currency Type

Category	%	n
USD	95.71%	9583
Pound Sterling	2.60%	261
Euro	1.52%	153
Kr (Swedish Krona)	0.11%	12
Dkk (Danish Krona)	0.029%	3
Total	100%	10012

All campaigns were collected from GoFundMe platform's healthcare category and divided into six categories: cancer, IVF (in vitro fertilization), leukemia, health insurance, lymphoma and surgery. Each event contains article/ story, raised amount, goal amount, date, currency type and number of donors (funder).

4.2 Results

Regression analysis is a one of the powerful statistical methods, which is allows examining the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. Used data was collected from medical-related campaign's data related information, which are the goal amount, donors (Independent variable) and raised amount (Dependent variable)'s information from GoFundMe platform. Collected data allows to us to measure the how these factors (medical campaign's goal amount and raised amount) influence each other. In regression analysis, R^2 is a statistical measure, which is close the data are to the fitted regression line. Also, it is known as the coefficient of determination. In general, R square was 0.873 as shown in **Table 6**. This result means that used data was high fitted.

Table 6. Summary of Statistics

Summary Statistics for Listings			**Regression**	
Variable	Mean	S.D.	Multiple R	0.896
Raised amount	8291.86	168.20	R^2	0.802
Goal amount	50733	2053.081	Adjusted R^2	0.605
Donors	78.04	139.64	S.E.	0.444
			Observations	10012

One-way ANOVA is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. That hypothesis-based test, which is meaning that aims to evaluate multiple mutually exclusive theories about study's data.

Table 7. The result of One-way ANOVA

	Coeff.	**S.E.**	**t**	***p***		**df**	**SS**	**MS**	**F**	**Sig. F**
Intercept	3. 685	0.0193	-1.924	0.000	Regression	2	65.650	32.825	10.392	0.331
Raised Amount	-3.875	8.861	-4.373	0.001	Residual	10010	32613.297	3.159		
Goal Amount	-1.093	8.665	-2.246	0.002	Total	10012	31678.947			

In addition, one-way ANOVA compares three or more categorical groups to determine whether there is a difference between the two groups. Each group must have at least three observations. This study considered raised amount, goal amount and number of donors. In **Table 7**, p-value is the probability of obtaining test results at least as extreme as the results observed during the test, assuming that the alternative hypothesis is correct. Thus, goal amount is affected to raised amount. If the p-value is less than 0.05, is statistically significance. In this

analysis, p-value was less than 0.05, which was 0.001 and 0.002. Thus, we continued the next step.

4.2.1 The Analysis through LDA (Study 1)

The purpose of LDA is helpful for words mainly used in healthcare-related fraud crowdfunding campaigns. The LDA of this study is based on two types. The first type is to determine the words mainly used for text information of medical-related crowdfunding campaigns, such as articles or stories about the campaign. The second type is to check what words are mainly written in very fraudulent behavioral medical crowdfunding campaigns such as campaigns by GoFundMe (Multiple different story campaigns from the same person). For example, Douglas Monahan and his company, iBackPack of Texas, LLC, used four crowdfunding campaigns to raise over $800,000. The FTC determined that most of the money was used for personal expenses

The third research question is, “are Latent Dirichlet allocation (LDA) and Collaborative Filtering (CF) effective in detecting fraud in crowdfunding campaigns?”. The LDA model was used to classify medical related crowdfunding campaigns related text into total of 10 topics.

Table 8. Keywords Order for 10 Topics

Order	Topic-1 (Resource for family)	Topic-2 (Medical bills)	Topic-3 (Wish)	Topic-4 (Financial assistance)	Topic-5 (Blood disease)	Topic-6 (Women)	Topic-7 (Incident)	Topi-8 (Transplantation)	Topic-9 (Recover)	Topic-10 (Patient)
1	support (0.038)	battle (0.033)	fund (0.045)	expense (0.056)	cancer (0.098)	IVF (0.131)	surgery (0.039)	cancer (0.046)	leukemia (0.038)	cancer (0.086)
2	family (0.033)	surgery (0.024)	baby (0.016)	medical (0.045)	fight (0.066)	journey (0.073)	lymphoma (0.026)	leukemia (0.030)	fund (0.032)	fight (0.064)
3	fund (0.022)	fund (0.021)	family (0.015)	help (0.037)	lymphoma (0.062)	baby (0.069)	Medical fund (0.024)	breast (0.020)	recovery (0.029)	treatment (0.019)
4	Health insurance (0.010)	family (0.019)	family (0.014)	leukemia (0.022)	Hodgkin (0.014)	journey (0.015)	fight (0.016)	transplant (0.019)	team (0.016)	battle (0.019)
5	fund (0.010)	help (0.013)	help (0.012)	need (0.022)	fighting (0.012)	miracle (0.007)	treatment (0.014)	life (0.015)	road (0.012)	lymphoma (0.017)
6	journey (0.008)	fund (0.012)	bill (0.011)	cost (0.011)	beat (0.006)	operation (0.006)	cancer (0.014)	stage (0.015)	memorial (0.012)	leukemia (0.016)
7	IVF (0.008)	cancer (0.011)	expense (0.006)	insurance (0.008)	brain (0.005)	leukemia (0.005)	helping (0.010)	fund (0.014)	beat (0.011)	mom (0.010)
8	battle (0.007)	family (0.011)	treatment (0.006)	battle (0.007)	Hodgkins (0.005)	dream (0.005)	family (0.008)	cell (0.007)	lymphoma (0.010)	support (0.008)
9	cancer (0.006)	expense (0.010)	cancer (0.005)	fund (0.007)	kick (0.004)	insurance (0.003)	Cancer fund (0.007)	bones (0.006)	cancer (0.009)	leukemia (0.008)
10	health (0.006)	help (0.006)	brain (0.005)	memory (0.006)	family (0.004)	dream (0.003)	brain (0.007)	kidney (0.005)	cancer (0.009)	surgery (0.006)

As shown in **Table 8**, words mainly used in a total of 10,012 medical-related campaign text information were characterized. In topic 1, keywords were associated with resource for family, topic 2 and 4, contained keywords implying the medical bills and financial assistance related keywords. Also, topic 6 and 10 included keywords on person-focused words. Furthermore, topic 5 focused on disease related words and topic 7, 8 and 9 was contained keywords that health related activities such as incident, transplantation, recover. Another topic 3 was focused

on wish. Next, we determined what words are mainly used for text information in healthcare-related crowdfunding campaigns, as shown in **Table 9**. It also explores and identifies a series of topics for six types of healthcare-related crowdfunding campaigns, including cancer, IVF (in vitro fertilization), leukemia, health insurance, lymphoma and surgery.

Table 9. The Centrality of Text-based Keywords in Six Campaigns Related to GoFundMe

Main Keywords	Topic-1 (Resource for family)	Topic-2 (Medical bills)	Topic-3 (Wish)	Topic-4 (Financial assistance)	Topic-5 (Blood disease)	Topic-6 (Women)	Topic- 7 (Incident)	Topic- 8 (Transplantation)	Topic- 9 (Recover)	Topic- 10 (Patient)
Cancer	0.195	0.010	0.516	0.550	0.093	0.185	0.132	0.033	0.004	0.014
IVF	0.147	0.029	0.139	0.087	0.009	0.001	0.066	0.882	0.012	0.019
Leukemia	0.023	0.029	0.045	0.090	0.186	0.506	0.037	0.033	0.442	0.112
Health Insurance	0.042	0.006	0.015	0.008	0.014	0.099	0.611	0.010	0.204	0.256
Lymphoma	0.009	0.454	0.138	0.171	0.054	0.065	0.047	0.012	0.020	0.448
Surgery	0.201	0.471	0.146	0.095	0.645	0.144	0.107	0.029	0.139	0.151

Table 10 provides subject-centric information for 10 types. To characterize and determine the keywords, the total data (total event number 10012) parsed text messages from 835 events that were repeated more than once. In **Table 12**, the keywords for Topics 1 and 7 are related to cancer care and cancer patient. Topic 2, 3, 4 and 5 contained more detailed keywords that suggest differences between keywords related to disease status. Topic 6 also contains basic explanatory keywords for disease stages. Topics 8 and 10 also show that keywords focus more on words related to the treatment process and common surgical procedure. Finally, the keywords for Topic 9 were mainly related to financial support.

Table 10. Keyword's List of Repeated Healthcare Campaign's Text Information on GoFundMe

Keyword Order (DC)	Topic-1 (Cancer care)	Topic-2 (Cardiology disease)	Topic-3 (Fertilization process)	Topic-4 (Blood disease)	Topic-5 (IVF fund)	Topic-6 Disease stage)	Topic-7 (Incident)	Topi-8 (Treatment)
1	home (0.061)	beat (0.055)	baby (0.054)	surgery (0.038)	baby (0.041)	parenthood (0.045)	cancer (0.065)	chemo (0.037)
2	fundraiser (0.058)	heart (0.051)	surgery (0.048)	health (0.034)	treatment (0.03)	health (0.041)	heart (0.054)	Non-Hodgkin (0.034)
3	kid (0.055)	heart (0.048)	need (0.045)	lymphoma (0.027)	surgery (0.0)	round (0.027)	life (0.043)	IV (0.031)
4	story (0.047)	friend (0.044)	support (0.039)	funeral (0.021)	IVF (0.023)	Medical fund (0.018)	kick (0.035)	bill (0.022)
5	healing (0.028)	year (0.041)	infertility (0.033)	support (0.019)	treatment (0.018)	tumor (0.011)	team (0.026)	memorial (0.014)
6	cancer (0.022)	home (0.035)	fighting (0.028)	surgery (0.013)	donation (0.015)	healing (0.009)	warrior (0.012)	baby (0.004)
7	battle (0.018)	lymphoma (0.023)	IVF (0.012)	funeral (0.007)	baby (0.012)	fight (0.005)	Cancer battle (0.008)	god (0.003)

8	heart (0.009)	cancer (0.017)	treatment (0.011)	mom (0.006)	hope (0.009)	heart (0.003)	memorial (0.006)	hope (0.001)
9	benefit (0.005)	mom (0.012)	life (0.004)	fund (0.003)	journey (0.002)	infertility (0.002)	fertility (0.003)	Cancer journey (0.004)
10	bill (0.002)	recovery (0.004)	adoption (0.004)	need (0.002)	ovarian (0.002)	stage (0.001)	prayer (0.003)	fight (0.003)

NOTE. DC: Degree Centrality

In summary, **Tables 8 and 10** show that there are differences in the use of words for crowdfunding text messages related to healthcare. Therefore, the use of LDA can help detect fraudulent healthcare-related crowdfunding activities in non-fraud healthcare-related crowdfunding campaigns.

4.2.2 The Analysis through Collaborative Filtering (Study 2)

The purpose of collaborative filtering (model-based approach) is to propose a reliable recommended model. It can calculate the difference between rating predictions for different collaborative filtering algorithm approaches. Therefore, when the performance of the recommendation system is effective, detection information for fraudulent healthcare campaigns may be presented. Therefore, the CF algorithm can be used to verify that the terms related to malpractice healthcare activities proposed in Study 1 are correct. The recommending algorithm used in this study identified activities containing possible fraudulent activity to donors (funders who donated to campaign creators) based on 9,177 healthcare-related crowdfunding activities (excluding 835 repeated campaigns). In model-based CF, SVD++ can be effective in detecting fraudulent content based on text data. The singular value decomposition (SVD)++ algorithm is employed as an optimized SVD algorithm to enhance the accuracy of prediction by generating implicit feedback [71]. However, the SVD++ algorithm is limited primarily by its low efficiency of calculation in the recommendation. To address this limitation of the algorithm, this study addressed social SVD++ algorithm. Social SVD++ considers the social network data as one of the implicit evaluation data. This algorithm is assumed that the preference of a certain user is related to the preference of the other user having socially connected relation [72]. Additionally, memory-based and model-based methods were compared by adding a user-to-user algorithm [58] to distinguish the difference between the recommended system algorithms.

As shown in the result of **Table 11**, the SVD++ algorithm had a correlation coefficient of 0.618 and the Social SVD++ had a correlation coefficient of 0.338. Likewise, user-to-user CF showed a very high correlation level with a correlation coefficient of 0.998. From this result, it was confirmed that fraudulent behavior information was included in the detection process of 6424 campaigns. In addition, 2,753 campaigns included non-fraud information with a correlation coefficient of 0.663 at 30% sparsity level, which is the best result of this study (Refer to Appendix A and B).

Table 11. Comparison of Performance by CF Algorithm

Accuracy / Method		# of instances	30%			50%			80%		
			MAE	RMSE	CC	MAE	RMSE	Correlation	MAE	RMSE	CC
TSS	(1)	2753	0.143	0.318	0.229	0.143	0.315	0.217	0.141	0.314	0.211
	(2)	2753	0.152	0.317	0.231	0.152	0.315	0.216	0.154	0.314	0.211

	(3)	2753	0.149	0.324	0.663	0.147	0.320	0.163	0.147	0.316	0.185
TRS	(1)	6424	0.125	0.274	0.618	0.124	0.277	0.583	0.137	0.298	0.311
	(2)	6424	0.147	0.300	0.337	0.143	0.298	0.352	0.147	0.293	0.366
	(3)	6424	0.002	0.022	0.998	0.002	0.021	0.998	0.000	0.007	0.990

NOTE. TSS: Test Set, TRS: Training Set, (1): SVD++, (2): Social SVD++, (3): User-to-User CF, CC: Correlation Coefficient

5. Conclusions

Crowdfunding platforms provide creators with the possibility to raise funds for projects or campaigns. Crowdfunding markets are largely inspired by the success of crowdfunding models over the Internet. However, as the number of platforms increases, so does the risk of fraud. In particular, the type of crowdfunding for health care (donation-based) is at greater risk of fraud because it relies on the complexity and responsibility of profit margins. This paper narrows the research gap by providing and evaluating methods for determining fraudulent healthcare crowdfunding behaviors, not healthcare-related crowdfunding behaviors.

The purpose of the new Hybrid Model of Family Detection Model (FDM) proposed in this study is to detect words mainly used in medical-related crowdfunding campaigns with very fraudulent intentions through the LDA method. In addition, LDA confirms the accuracy of the proposed fraud-related words and calculates and evaluates the differences between collaborative filtering methods such as SVD++, social SVD+++ and user-to-user CF.

The theoretical implications of this study are as follows. As the number of platforms in the crowdfunding market increases, it is gaining popularity in creating campaigns, posting project ideas, and raising funds. Behind it leads to the risk of fraud. Expanding previous studies that mainly consider questions about factors or factors leading to successful funding of campaigns or projects, this study dealt with how to determine fraudulent behavior in campaigns (Research Question 1) rather than fraud using a detection approach through technical applications. Second, the results of this study confirmed that fraud detection in the field of medical crowdfunding can be appropriately solved by the proposed recommendation system. Therefore, evidence was provided that analysis and detection of fraudulent content should be performed using LDA and collaborative filtering to detect fraudulent behavior of crowdfunding-related platforms (Research Question 1). Third, since collaborative filtering with LDA provides good classification results, it is valuable for fraud detection (Research Questions 2 and 3). As can be seen from the evaluation results of machine learning and regression analysis, the hybrid model integrated into the FDM (Fraud Detection Model) provides optimal detection based on text information from healthcare-related crowdfunding campaigns.

The practical implications of this study are as follows: First, from a practical point of view, it has a lot to do with various donors on the fundraising platform. Potential campaign donors can reduce the risk of fraud by detecting an approach to fraud. Second, efforts should be made from a corporate perspective to ensure the integrity of crowdfunding platforms and prevent fraud. Therefore, the proposed malpractice detection model is also beneficial for platform operations.

To identify fraud cases, this study is limited to: This study is campaigns for the GoFundMe platform, a donation-based platform. If more fraud is detected in the crowdfunding market, it should be other types of crowdfunding, such as reward, loans, or equity-based crowdfunding. Nevertheless, this study found that the social capital of crowdfunding-related beneficiaries (e.g., the number of Facebook friends) was related to the success of the campaign when used

with comments. In donation-based campaigns, donors mainly used commentary functions. In addition, this study uses limited information such as articles, stories, dates, target amounts, amounts raised, and the number of donors. Therefore, in future studies, it is expected that a wide range of topics, such as social capital information of beneficiaries (campaign creators), will be able to identify more diverse fraud detection processes.

References

[1] M. Siering, J.-A. Koch and A. V. Deokar, "Detecting Fraudulent Behavior on Crowdfunding Platforms: The Role of Linguistic and Content-Based Cues in Static and Dynamic Contexts," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 421-455, 2016. Article (CrossRef Link)

[2] A. Abbasi, F. M. Zahedi, D. Zeng, Y. Chen, H. Chen and J. F. Nunamaker, "Enhancing Predictive Analytics for Anti-Phishing by Exploiting Website Genre Information," *Journal of Management Information Systems*, vol. 31, no. 4, pp. 109-157, 2015. Article (CrossRef Link)

[3] L. Zhou, J. K. Burgoon, J. F. Nunamaker and D. Twitchell, "Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications," *Group Decision and Negotiation*, vol. 13, no. 1, pp. 81-106, 2004. Article (CrossRef Link)

[4] L. Zhou and D. Zhang, "Following linguistic footprints," *Communications of the ACM*, vol. 51, no. 9, pp. 119-122, 2008. Article (CrossRef Link)

[5] J.-A. Koch and M. Siering, "Crowdfunding Success Factors: The Characteristics of Successfully Funded Projects on Crowdfunding Platforms," in *Proc. of the 23rd European Conference on Information Systems*, 2015. Article (CrossRef Link)

[6] B. Jin, H. Zhao, E. Chen, Q. Liu and Y. Ge., "Estimating the Days to Success of Campaigns in Crowdfunding: A Deep Survival Perspective," in *Proc. of AAAI*, 2019. Article (CrossRef Link)

[7] G. Burtch and J. Chan, "Investigating the Relationship Between Medical Crowdfunding and Personal Bankruptcy in the United States: Evidence of a Digital Divide," *MIS Quarterly*, vol. 43, no. 1, pp. 237-262, 2019. Article (CrossRef Link)

[8] J. G. Kim, H. K. Kong, K. Karahalios, W.-T. Fu and H. Hong, "The Power of Collective Endorsements," in *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4538-4549, 2016. Article (CrossRef Link)

[9] O. Sorenson, V. Assenova, G. C. Li, J. Boada and L. Fleming, "Expand innovation finance via crowdfunding," *Science*, vol. 354, no. 6319, pp. 1526-1528, 2016. Article (CrossRef Link)

[10] E. Gerber, J. Hui and P.-Y. Kuo, "Crowdfunding: Why People are Motivated to Post and Fund Projects on Crowdfunding Platforms," in *Proc. of Computer Supported Cooperative Work*, 2011. Article (CrossRef Link)

[11] G. Bassani, N. Marinelli and S. Vismara, "Crowdfunding in healthcare," *The Journal of Technology Transfer*, vol. 44, no. 4, pp. 1290-1310, 2019. Article (CrossRef Link)

[12] E. Mollick, "The dynamics of crowdfunding: An exploratory study," *Journal of Business Venturing*, vol. 29, no. 1, pp. 1-16, 2014. Article (CrossRef Link)

[13] P. Younkin and K. Kashkooli, "What Problems Does Crowdfunding Solve?," *Calif. Manage. Rev.*, vol. 58, pp. 20-43, 2016.

[14] C. W. Callaghan, "Crowdfunding To Generate Crowdsourced R&D: The Alternative Paradigm Of Societal Problem Solving Offered By Second Generation Innovation And R&D," *International Business& Economics Research Journal*, vol. 13, no. 6, pp. 1499-1514, 2014. Article (CrossRef Link)

[15] J. Snyder, P. Chow-White, V. A. Crooks and A. Mathers, "Widening the gap: additional concerns with crowdfunding in health care," *Lancet Oncol*, vol. 18, no. 5, pp. e240, 2017. Article (CrossRef Link)

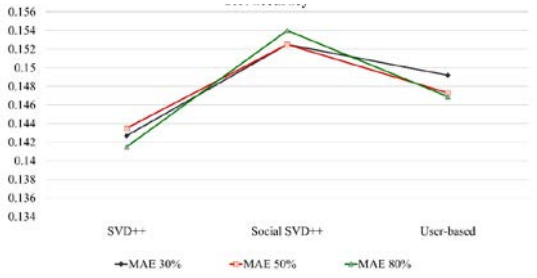
[16] D. C. Fumagalli and A. M. Gouw, "Crowdfunding for Personalized Medicine Research," *Yale J Biol Med*, vol. 88, no. 4, pp. 413-414, 2015. Article (CrossRef Link)

[17] G. Dressler and S. A. Kelly, "Ethical implications of medical crowdfunding: the case of Charlie Gard," *J Med Ethics*, vol. 44, no. 7, pp. 453-457, 2018. Article (CrossRef Link)
[18] G. Burtch, A. Ghose and S. Wattal, "An Empirical Examination of Peer Referrals in Online Crowdfunding," in *Proc. of International Confernece on Information Systems*, 2014. Article (CrossRef Link)
[19] G. Burtch, A. Ghose and S. Wattal, "An Empirical Examination of the Antecedents and Consequences of Contribution Patterns in Crowd-Funded Markets," *Information Systems Research*, vol. 24, no. 3, pp. 499-519, 2013. Article (CrossRef Link)
[20] P. Belleflamme, T. Lambert and A. Schwienbacher, "Individual crowdfunding practices," *Venture Capital*, vol. 15, no. 4, pp. 313-333, 2013. Article (CrossRef Link)
[21] J. V. d. Cruz, "Competition and Regulation of Crowdfunding Platforms: A Two-Sided Market Approach," *Communications & Strategies*, vol. 99, no. pp. 33-50, 2016. Article (CrossRef Link)
[22] A. Agrawal, C. Catalini and A. Goldfarb, "Some Simple Economics of Crowdfunding," *Innovation Policy and the Economy*, vol. 14, no. pp. 63-97, 2014. Article (CrossRef Link)
[23] T. Mitra and E. Gilbert, "The language that gets people to give: phrases that predict success on kickstarter," in *Proc. of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 49-61, 2014. Article (CrossRef Link)
[24] S. Pitschner and S. Pitschner-Finn, "Non-profit differentials in crowd-based financing: Evidence from 50,000 campaigns," *Econ. Letters*, vol. 123, no. 3, pp. 391-394, 2014. Article (CrossRef Link)
[25] K.-y. R. Choy and D. Schlagwein, "IT Affordances and Donor Motivations in ChariTable Crowdfunding: The "Earthship Kapita" Case," *Information Technology & People*, vol. 29, no. 1, 2016. Article (CrossRef Link)
[26] D. Wang, B. Chen and J. Chen, "Credit card fraud detection strategies with consumer incentives," *Omega*, vol. 88, no. pp. 179-195, 2019. Article (CrossRef Link)
[27] R. Rambola, P. Varshney and P. Vishwakarma, "Data Mining Techniques for Fraud Detection in Banking Sector," in *Proc. of 2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 14-15 Dec. 2018. Article (CrossRef Link)
[28] M. Kirlidog and C. Asuk, "A Fraud Detection Approach with Data Mining in Health Insurance," *Procedia - Social and Behavioral Sciences*, vol. 62, no. pp. 989-994, 2012. Article (CrossRef Link)
[29] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585-594, 2011. Article (CrossRef Link)
[30] F. Zahedi, A. Abbasi and Y. Chen, "Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance," *J. Assoc. Inf. Syst.*, vol. 16, pp. 2, 2015. Article (CrossRef Link)
[31] H. Landström, A. Parhankangas and C. Mason, *Handbook of Research on Crowdfunding*, Cheltenham, UK, Edward Elgar Publishing, 2019
[32] J. R. Carlson and R. W. Zmud, "Channel Expansion Theory and the Experiential Nature of Media Richness Perceptions," *Academy of Management Journal*, vol. 42, no. 2, pp. 153-170, 1999. Article (CrossRef Link)
[33] R. L. Daft and R. H. Lengel, "Organizational Information Requirements, Media Richness and Structural Design," *Management Science*, vol. 32, no. 5, pp. 554-571, 1986. Article (CrossRef Link)
[34] P. Johnson, "Detecting deception: adversarial problem solving in a low base-rate world," *Cognitive Science*, vol. 25, no. 3, pp. 355-392, 2001. Article (CrossRef Link)
[35] P. E. Johnson, S. Grazioli and K. Jamal, "Fraud detection: Intentionality and deception in cognition," *Accounting, Organizations and Society*, vol. 18, no. 5, pp. 467-488, 1993. Article (CrossRef Link)
[36] A. Vrij, "Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies," *Psychol. Pub. Pol'y & L.*, vol. 11, no. 1, pp. 3-41, 2005. Article (CrossRef Link)
[37] M. Zuckerman, B. M. DePaulo and R. Rosenthal, "Verbal and Nonverbal Communication of Deception," *Advances in Experimental Social Psychology*, vol. 14, pp. 1-59, 1981. Article (CrossRef Link)

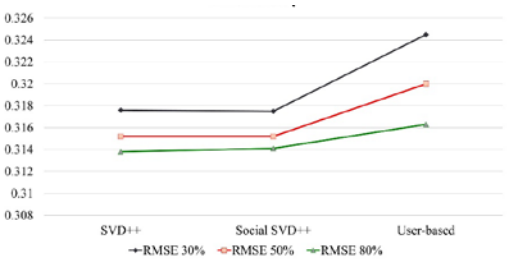
[38] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88-106, 1969. Article (CrossRef Link)
[39] S. A. McCornack, "Information manipulation theory," *Communication Monographs*, vol. 59, no. 1, pp. 1-16, 1992. Article (CrossRef Link)
[40] J. K. Burgoon, L. Dillman and L. A. Stem, "Adaptation in Dyadic Interaction: Defining and Operationalizing Patterns of Reciprocity and Compensation," *Communication Theory*, vol. 3, no. 4, pp. 295-316, 1993. Article (CrossRef Link)
[41] D. B. Buller, J. K. Burgoon, A. Buslig and J. Roiger, "Testing Interpersonal Deception Theory: The Language of Interpersonal Deception," *Communication Theory*, vol. 6, no. 3, pp. 268-289, 1996. Article (CrossRef Link)
[42] L. N. Driscoll, "A Validity Assessment of Written Statements from Suspects in Criminal Investigations Using the Scan Technique," *Police Studies*, vol. 17, no. 4, pp. 77-88, 1994. Article (CrossRef Link)
[43] D. Xing and M. Girolami, "Employing Latent Dirichlet Allocation for fraud detection in telecommunications," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1727-1734, 2007. Article (CrossRef Link)
[44] J. Zhang, M. Marszałek, S. Lazebnik and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213-238, 2007. Article (CrossRef Link)
[45] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman and W. T. Freeman, "Discovering object categories in image collections," M. I. o. Technology, 2005. Article (CrossRef Link)
[46] G. Csurka, "Visual categorization with bags of keypoints," in *Proc. of eccv*, 2004. Article (CrossRef Link)
[47] Sivic and Zisserman, "Video Google: a text retrieval approach to object matching in videos," in P*roc. of Ninth IEEE International Conference on Computer Vision*, 13-16 Oct. 2003. Article (CrossRef Link)
[48] H. Wang and E. Overby, "How Does Online Lending Influence Bankruptcy Filings? Evidence from a Natural Experiment," *Academy of Management Proceedings*, vol. 2017, no. 1, pp. 15937, 2017. Article (CrossRef Link)
[49] P. Seemakurthi, S. Zhang and Y. Qi, "Detection of fraudulent financial reports with machine learning techniques," in *Proc. of 2015 Systems and Information Engineering Design Symposium*, 24-24 April 2015. Article (CrossRef Link)
[50] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *Proc. of 2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, 15-17 Jan. 2015. Article (CrossRef Link)
[51] K. D. Lee, K. Han and S.-H. Myaeng, "Capturing Word Choice Patterns with LDA for Fake Review Detection in Sentiment Analysis," in *Proc. of the 6th International Conference on Web Intelligence, Mining and Semantics*, pp. 1-7, 2016. Article (CrossRef Link)
[52] J. Li, C. Cardie and S. Li, "TopicSpam: a Topic-Model based approach for spam detection," in *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217-221, 2013. Article (CrossRef Link)
[53] V. Sandulescu and M. Ester, "Detecting Singleton Review Spammers Using Semantic Similarity," in *Proc. of the 24th International Conference on World Wide Web*, pp. 971-976, 2015. Article (CrossRef Link)
[54] F. Lyu and J. Choi, "The Forecasting Sales Volume and Satisfaction of Organic Products through Text Mining on Web Customer Reviews," *Sustainability*, vol. 12, no. 11, 2020. Article (CrossRef Link)
[55] X. He, L. Liao, H. Zhang, L. Nie, X. Hu and T.-S. Chua, "Neural Collaborative Filtering," in *Proc. of the 26th International Conference on World Wide Web*, pp. 173-182, 2017. Article (CrossRef Link)
[56] A. Beutel, K. Murray, C. Faloutsos and A. J. Smola, "CoBaFi: collaborative bayesian filtering," in *Proc. of the 23rd international conference on World wide web - WWW '14*, pp. 97-108, 2014. Article (CrossRef Link)

[57] R. Hu and P. Pu, "Acceptance issues of personality-based recommender systems," in *Proc. of the third ACM conference on Recommender systems - RecSys '09*, pp. 221-224, 2009. Article (CrossRef Link)
[58] J. Choi, H. J. Lee and Y. C. Kim, "The Influence of Social Presence on Customer Intention to Reuse Online Recommender Systems: The Roles of Personalization and Product Type," *International Journal of Electronic Commerce*, vol. 16, no. 1, pp. 129-154, 2011. Article (CrossRef Link)
[59] B. Pathak, R. Garfinkel, R. D. Gopal, R. Venkatesan and F. Yin, "Empirical Analysis of the Impact of Recommender Systems on Sales," *Journal of Management Information Systems*, vol. 27, no. 2, pp. 159-188, 2010. Article (CrossRef Link)
[60] J. Choi, H. J. Lee, F. Sajjad and H. Lee, "The influence of national culture on the attitude towards mobile recommender systems," *Technol. Forecast. Soc. Change*, vol. 86, no. pp. 65-79, 2014. Article (CrossRef Link)
[61] J. Choi, H. J. Lee and H.-W. Kim, "Examining the effects of personalized App recommender systems on purchase intention: A self and social-interaction perspective," *Journal of Electronic Commerce Research*, vol. 18, no. 1, pp. 73-102, 2017. Article (CrossRef Link)
[62] L. Chen, F. Hsu, M. Chen and Y. Hsu, "Developing recommender systems with the consideration of product profitability for sellers," *Information Sciences*, vol. 178, no. 4, pp. 1032-1048, 2008. Article (CrossRef Link)
[63] M. Jalali, N. Mustapha, M. N. Sulaiman and A. Mamat, "WebPUM: A Web-based recommendation system to predict user future movements," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6201-6212, 2010. Article (CrossRef Link)
[64] J. Han, J. Pei and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2011
[65] H. Lieberman, "Letizia: an agent that assists web browsing," in *Proc. of the 14th international joint conference on Artificial intelligence - Volume 1*. [Online]. Available: https://web.media.mit.edu/~lieber/Lieberary/Letizia/Letizia-AAAI/Letizia.html
[66] M. J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering," *Artificial Intelligence Review*, vol. 13, no. 5/6, pp. 393-408, 1999. Article (CrossRef Link)
[67] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005. Article (CrossRef Link)
[68] M. Göksedef and Ş. Gündüz-Öğüdücü, "Combination of Web page recommender systems," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 2911-2922, 2010. Article (CrossRef Link)
[69] M. B. Meghani and M. Mulchandani, "A survey of Anomaly Detection methods in networks and Collaborative Filtering Recommender Systems," 2016. Article (CrossRef Link)
[70] M.-P. T. Do, D. V. Nguyen and L. Nguyen, "Model-based Approach for Collaborative Filtering," in *Proc. of The 6th International Conference on Information Technology for Education*, pp. 217-228, 2010. Article (CrossRef Link)
[71] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426-434, 2008. Article (CrossRef Link)
[72] R. Kumar, B. Verma and S. Rastogi, "Social Popularity based SVD++ Recommender System," *International Journal of Computer Applications*, vol. 87, no.14, pp. 33-37, 2014. Article (CrossRef Link)

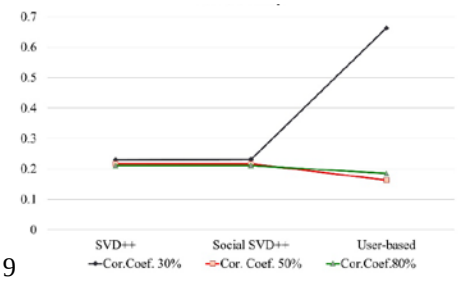
APPENDIX A. Accuracy-based on (1) MAE, (2) RMSE, (3) Correlation Coefficient levels from Test sets



(1)



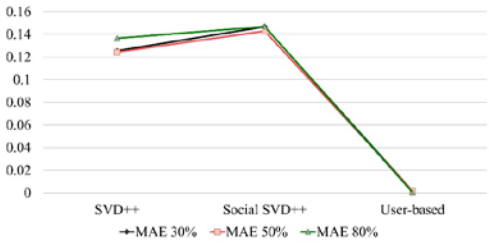
(2)



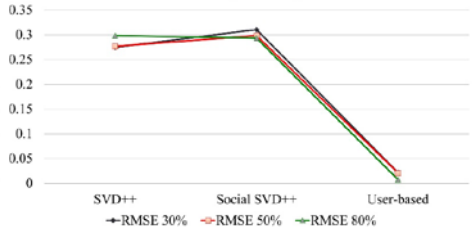
9

(3)

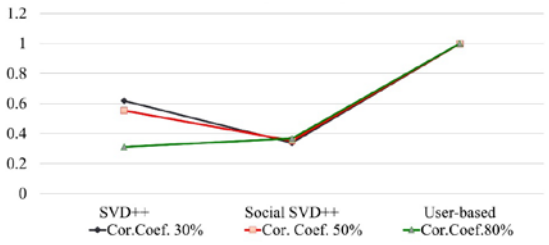
APPENDIX B. Accuracy-based on (1) MAE, (2) RMSE, (3) Correlation Coefficient levels from Training sets



(1)



(2)



(3)

Jaewon Choi is an associate professor and the head of the Department of Business Administration, Global Business School, Soonchunhyang University. His research areas are bigdata analytics, artificial intelligence, social network analysis, block chain, text mining, personalized intelligent agents in e-commerce and m-commerce. He published papers on *Journal of Electronic Commerce Research*, *International Journal of Electronic Commerce*, *Technical Forecasting and Social Change*, *Cyberpsychology Behavior and Social Networking*, and other journals.

Professor Jaehyoun Kim received his B.S. degree in mathematics from Sungkyunkwan University, Seoul, Korea, M.S. degree in computer science from Western Illinois University and Ph.D. degrees in computer science from Illinois Institute of Technology in U.S.A. He was a Chief Technology Officer at Kookmin Bank in Korea before he joined the Department of Computer Education at Sungkyunkwan University(SKKU) in March 2002. Currently he is a professor at Sungkyunkwan University. Also, he is a dean of College of Education and a chairman of Data Science (DS) Education Center at SKKU. His research interests include software engineering & architecture, e-Learning, SW/AI education and computer based learning.

Ho Lee is an assistant professor in the department of Future Technology at Korea University of Technology and Education. He completed a Ph.D. in Information Systems at Yonsei University, Korea and received his Bachelor of Science in Computer Science from State University of New York at Sony Brook, USA. His current research interests are in the areas of anonymity, online behavior, knowledge management, job change and data analytics.