applied sciences

MDPI

Article

Topic Predictions and Optimized Recommendation Mechanism Based on Integrated Topic Modeling and Deep Neural Networks in Crowdfunding Platforms

Wafa Shafqat and Yung-Cheol Byun *

Department of Computer Engineering, Jeju National University, Jeju 63243, Korea; wafashafqat92@gmail.com
* Correspondence: ycb@jejunu.ac.kr

Received: 18 October 2019; Accepted: 6 December 2019; Published: 13 December 2019

Abstract: The accelerated growth rate of internet users and its applications, primarily e-business, has accustomed people to write their comments and reviews about the product they received. These reviews are remarkably competent to shape customers' decisions. However, in crowdfunding, where investors finance innovative ideas in exchange for some rewards or products, the comments of investors are often ignored. These comments can play a markedly significant role in helping crowdfunding platforms to battle against the bitter challenge of fraudulent activities. We take advantage of the language modeling techniques and aim to merge them with neural networks to identify some hidden discussion patterns in the comments. Our objective is to design a language modeling based neural network architecture, where Recurrent Neural Networks (RNN) Long Short-Term Memory (LSTM) is used to predict discussion trends, i.e., either towards scam or non-scam. LSTM layers are fed with latent topic distribution learned from the pre-trained Latent Dirichlet Allocation (LDA) model. In order to optimize the recommendations, we used Particle Swarm Optimization (PSO) as a baseline algorithm. This module helps investors find secure projects to invest in (with the highest chances of delivery) within their preferred categories. We used prediction accuracy, an optimal number of identified topics, and the number of epochs, as metrics of performance evaluation for the proposed approach. We compared our results with simple Neural Networks (NNs) and NN-LDA based on these performance metrics. The strengths of both integrated models suggest that the proposed model can play a substantial role in a better understanding of crowdfunding comments.

Keywords: topic modeling; Latent Dirichlet Allocation (LDA); Recurrent Neural Networks (RNN); Long Short-Term Memory (LSTM); optimization; Particle Swarm Optimization (PSO); crowdfunding scams; Kickstarter; objective function

1. Introduction

With the radical evolution in information technology and the growing number of internet users (approximately 4 billion internet users across the globe in 2018) [1]), internet usage has turned out to be a significant communication bridge between customers and organizations. Different social networking sites or discussion portals enable this communication by providing a facility to share customers' opinions and reviews. These reviews are wealthy of information such as sentiments, critics, and customers' concerns [2]. The in-depth analysis of these reviews is helpful in different applications, such as predicting trends, discussion topics, and popular keywords.

Though the easy accessibility of the internet is a blessing, it brings many challenges too. The increasing number of internet misuse cases in the form of fraud, harassment, and information leakage has become a prevalent concern for the users and administrative parties.

According to the Internet Crime Report 2018 [3], there are billions of dollars recorded as fraud on the internet, as shown in Figure 1. Crowd-thieving is enabling fraudsters to illegally raise money in terms of misrepresenting their ideas, taking an advance fee, investments, non-payments, or non-delivery or personal data breach, as shown in Figure 2.

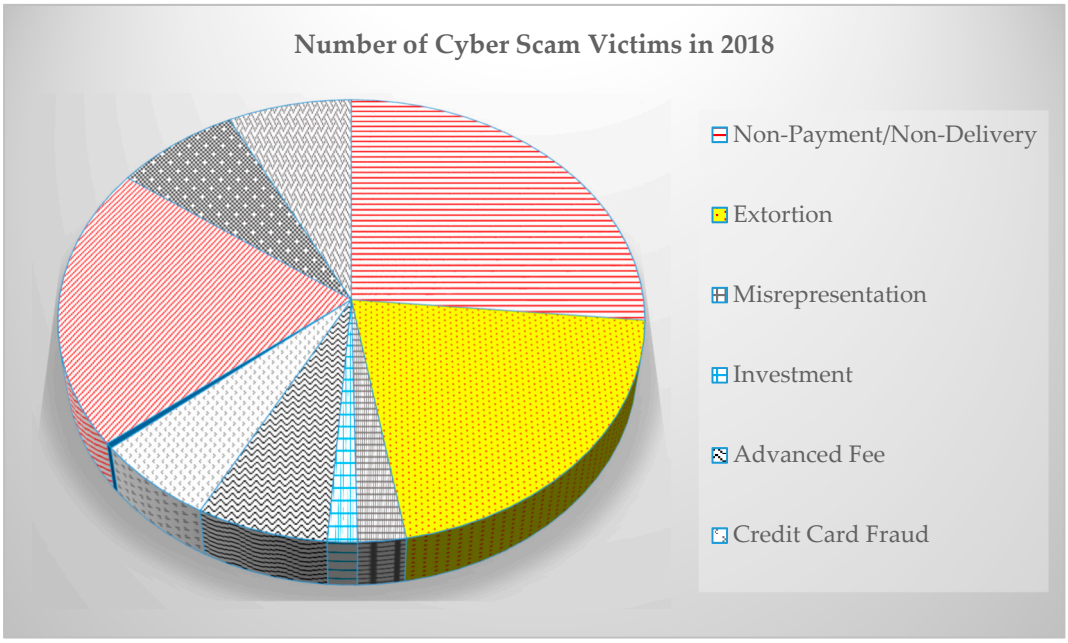

Figure 1. The victims of cyber scams according to the Internet Crime Report 2018.

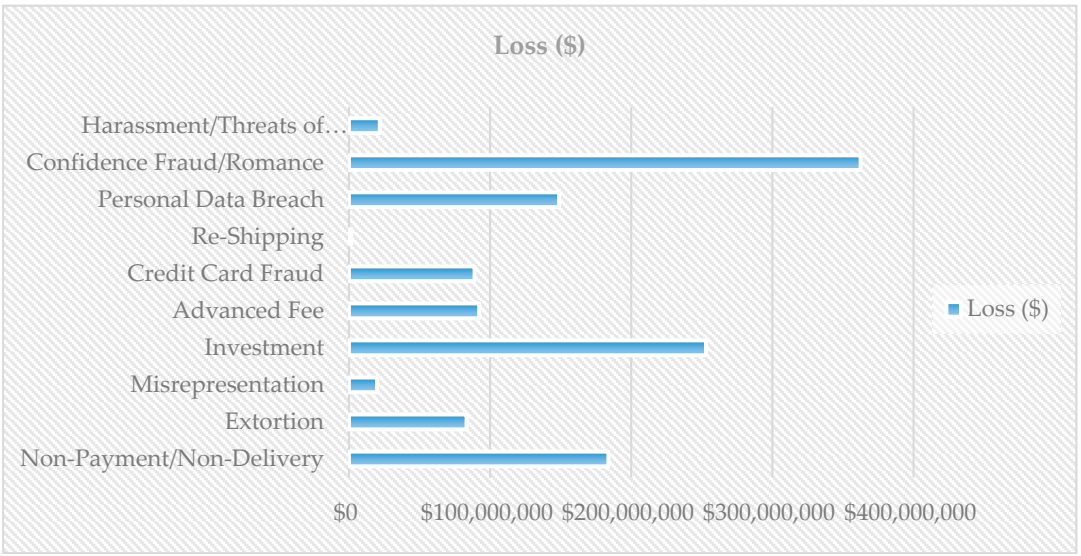

Figure 2. Money lost due to cyber scams according to the Internet Crime Report 2018.

Crowdfunding has emerged to be a modern and provocative process to attract a substantial number of investors towards innovative startups. It also withstands the challenges of ensuring accountability, regulating laws, administrating ethics, and handling funds [4]. The idea that anyone with an innovative concept can start to collect funds for a product, and the arduous process to access a campaign's legitimacy [5], raise a growing sense of alarm for the crowdfunding platforms. There is a dire need to build a recommendation system to suggest reliable projects to investors according to their preferences. It is essential to have a reasonable amount of data for these recommendation systems to perform well. To address this challenge, utilizing comments or reviews is the right approach as these

reviews represents the user preferences and product characteristics. Therefore, latent feature vectors can be obtained comparatively in a more natural way. In previous studies, user reviews are analyzed primarily through topic modeling-based approaches such as Latent Dirichlet Allocation (LDA), which aims to discover hidden themes in textual documents [6]. However, topic modeling approaches fail to capture contextual information [7].

In comparison with topic modeling, deep learning approaches, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), can capture the contextual and temporal dependencies. Regardless of the strength and power of these deep learning methods, there are still some limitations in terms of the sliding window size. For example, if the window size is too small, it can cause CNN to fail in connecting words to sentences. Similarly, RNN finds it challenging when it comes to sentence size; in other words, if a review is too long, RNN's ability to extract contextual information will be limited.

To alleviate the problem, we suggest the integration of deep learning approaches with language modeling techniques. Therefore, we propose a hybrid method that takes advantage of both the solution domains, to provide a deeper understanding of the comments. The deep learning methods can help reserve the context information, whereas topic modeling focuses on word relationships, which mollifies the chances of information loss. For the language modeling part, we use the most common topic modeling approach, i.e., LDA. For deep learning, we opted for a particular RNN type known as Long-Short-Term Memory (LSTM) network because of its strengths of reserving time-dependent information and improved performance over other neural systems. Finally, these two modules are integrated into a hybrid framework resulting in a better model towards topic clustering and predictions. The results of this module can be utilized in other applications such as recommendations based on user preferences, or sentiment classification.

To make some project recommendations for the investors, we have built a recommendation system on top of the prediction module, which is based on the best predictions and user preferences. In order to make an ideal recommendation system for investors and to save them from the risks of potential frauds, extensive contextual learning is required. In summary, the proposed system based on a hybrid approach primarily consists of two major modules: (1) Prediction module (2) Optimized recommendation module. In the prediction module, the latent topics are learned through LDA and are fed to LSTM. LSTM then predicts the topic class for the next word in the corpus. The optimal recommendation module takes the user's preferences into account in combination with the projects having the highest level of authenticity. The authenticity level (a measure of a project's credibility) is measured on the bases of the topic clusters. If topic clusters fall into topics depicting suspicions, then authenticity level will be lower. The user preferences include project category, project location, level of rewards, and delivery date. We have also formulated an objective function based on Particle Swarm Optimization (PSO), which aims to find optimal recommendations for the user for ongoing crowdfunding projects. The objective function maximizes the user preferences and minimizes the impact of topic classes, which have low authenticity levels.

The rest of the paper is divided as follows. Section 2 presents the literature review related to our proposed mechanism, Section 3 presents the algorithmic approaches such as basics of language modeling, LSTMs, and PSO. Section 4 presents the propose model in detail. Section 5 sheds light on the implementation and experimental environment. Section 6 presents the preliminary results and a comparative analysis of the proposed approach with baseline algorithms. In last section, we discuss the challenges, limitation, and contributions of our study.

2. Related Works

In this section, we present literature review of the related works. Section 2.1 presents the related works on crowdfunding dynamics, Section 2.2 presents the predictions and recommendations in crowdfunding, and Section 2.3 briefly presents related works on topic modeling and RNNs.

2.1. Crowdfunding Dynamics

As our study focuses on project class prediction with timely and reliable recommendations, it is crucial to analyze the effect of different factors at an early stage of the project. Communication, in the form of updates, is used as a key to increase investments during the funding period by the project creators in equity-based crowdfunding [8]. Some studies focus on investigating the dynamics of reward-based crowdfunding [9–13]. In [14], a field experiment was performed to study the effect of an investment in a project which was previously unfunded. In a more recent study [15], it has shown that the high-profile investors' actions, at an early stage of the project funding, create a cascading effect, and other investors learn from them and follow their steps which may lead to successful funding of a project. This cascading effect can also lead to unwanted outcomes if the uncertainty is high or the information gap is high [16]. According to the discussions in [10], successful campaigns during their starting period of fundraising were found to have relatively significant investments. In addition to that, the experiments are performed on Kickstarter data, where several investors were observed over time and results show that investors' support increases as the project enter the final stage of the funding. One of our motivations for this study is the rising number of studies on reward-based crowdfunding [17,18] and equity crowdfunding [19,20].

2.2. Predictions and Recommendations in Crowdfunding

There are various studies on crowdfunding that target to predict different trends and project success [21–24]. In [22], a tool is built to get reviews on their project ideas for a startup. Some studies have explored different linguistic features, specific patterns, and writing styles of project creators to reveal the impact of language on the success or failure of a campaign [25–27].

In [28], crowdfunding success prediction is estimated through a text analytics approach, where LDA is used to extract semantic features out of the text, along with feature selection, and data mining. In a similar work [29], crowdfunding updates are analyzed by using LDA to classify the updates into different topic categories.

2.3. Topic Models and RNNs

To capture the semantic features of the text, topic models play a vital role. The semantic features are extracted as latent topics. There have been numerous studies and applications of topic models since this idea was first introduced by Blei et al. [6]. These studies [30–32] have covered many research areas ranging from scientific studies to mathematical equations. Recently, many studies are using deep learning methods in combination with topic models. In [32], RNN is used along with topic models to enhance the performance of topic modeling for scientific texts. In [33,34], a neural topic model has been presented. Some recent studies have paid more attention and focused on neural variational inferences in order to train the topic models [35,36]. Moreover, for sequences that have long term dependencies, RNN proved to be an effective solution [37,38]. Other studies that have used RNNs to model different language-related problems include handwriting recognition [39], LATEX modeling [40], and semantic parsing [41], etc.

Our model is driven from a hybrid of the approaches which are based on a joint topic language model. The motivation behind these architectures is to extract the latent topics through topic models, e.g., LDA and these pre-trained models are used for Deep neural networks, such as RNNs or CNNs, modeling. In [42], a pre-trained LDA model was incorporated into an RNN model. In other studies [43,44], both topic models and neural networks are trained together, which we also aim to perform. A recent work [45] has proposed a Sentence Level Recurrent Topic Model (SLRTM), where for each sentence, a topic is decided based on a non-sequential Dirichlet structure similar to LDA. Therefore, it is not much useful for capturing long-range temporal dependencies, also as it uses the whole vocabulary, which enriches it with lots of features. In our case, we are using only topics, not the entire dictionary. There are other models introduced in [46,47], which are based on recurrent latent variables; in these models, RNN is enabled with latent variables in order to cater to the inconsistencies

in the input data. Though, these models focus on images and speech data; while we use a discrete input data space and uses text data and numeric data.

In the previous studies, LDA and deep learning models are combined to either improve the language modeling or on the same word text. In our case, we are building a recommendation model on top of LDA-LSTM hybrid model. User preferences are added as an extra layer of input. One key difference of our model is that the latent topics are used in the LSTM layer while other models, such as [39] and [43], integrate them in the output layer of LSTM.

3. Algorithmic Approaches used in the Methodology

In this section, we present the algorithmic approaches used in the proposed mechanism. Section 3.1 presents the related works on language modeling and LDA. Section 3.2 presents the related works on LSTM, and Section 3.3 presents related works on optimization algorithms.

3.1. Language Modeling

Language modeling depends on a function learning that calculates the log probability of an activity as $\log p(\omega|\text{model})$, or a sentence as $\omega = (\omega 1, \omega 2, \ldots, \omega n)$. This function is then used for the prediction of the next word or activity. It can also be used in different other ways, e.g., LDA uses a bag-of-words approach. Alternatively, it can be used in RNN modeling to prevent temporal information loss, to model $\log p(\omega n|\ \omega 1, \omega 2, \ldots, \omega n\text{-}1, \text{model})$.

Latent Dirichlet Allocation (LDA)

In our proposed system, LDA is used for topic modeling, Figure 3 presents the basic block diagram of LDA, where a user comment is treated as a document and fed to the LDA model, after some preprocessing. LDA results into clusters of similar words indicating a topic or theme. This process is repeated for all the comments in the dataset.

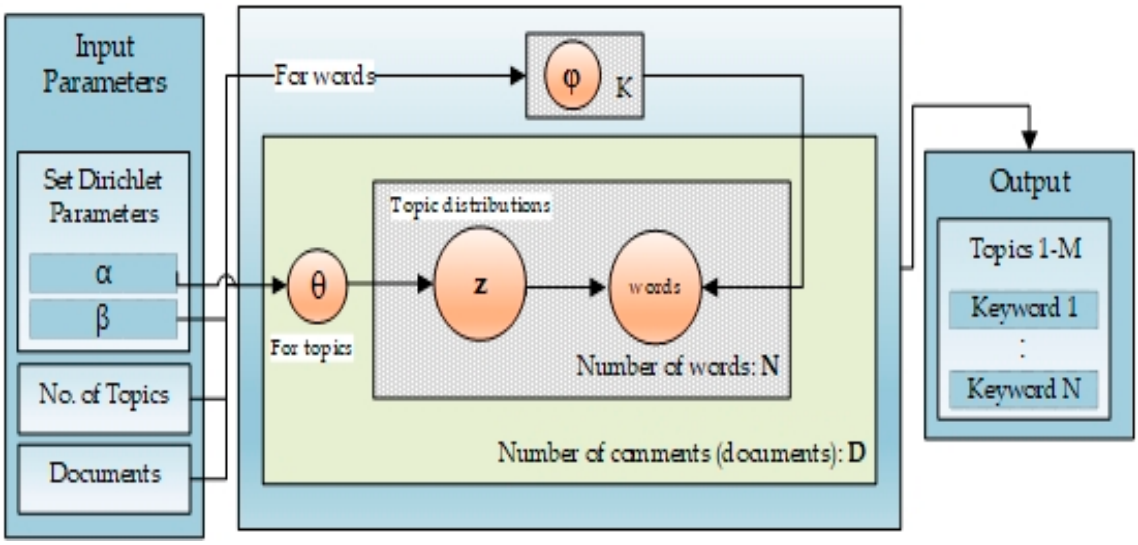

Figure 3. Plate Notation of Latent Dirichlet Allocation (LDA).

3.2. Long Short-Term Memory (LSTM)

The traditional topic modeling approaches have some limitations when it comes to context learning. For any language model, temporal aspects are also fundamental. Therefore, a well suitable method is required to perform this task. An RNN type, LSTM, can effectively learn the context and temporal features and can better classify or predict, primarily when we have large data sets with time-series information.

RNNs are the type of networks that generate recurrent connections to memorize. Language models based on recurrent neural networks have lately established state-of-the-art performance in different applications. The dynamic temporal behavior of RNN makes them favorable for sequential classification-based problems.

For training, it takes the first-word w from the sequence of input; the output h0 along with the next word w1 is taken as input in the next step, and so on. This way, it keeps remembering the context while training, as shown in Figure 4.

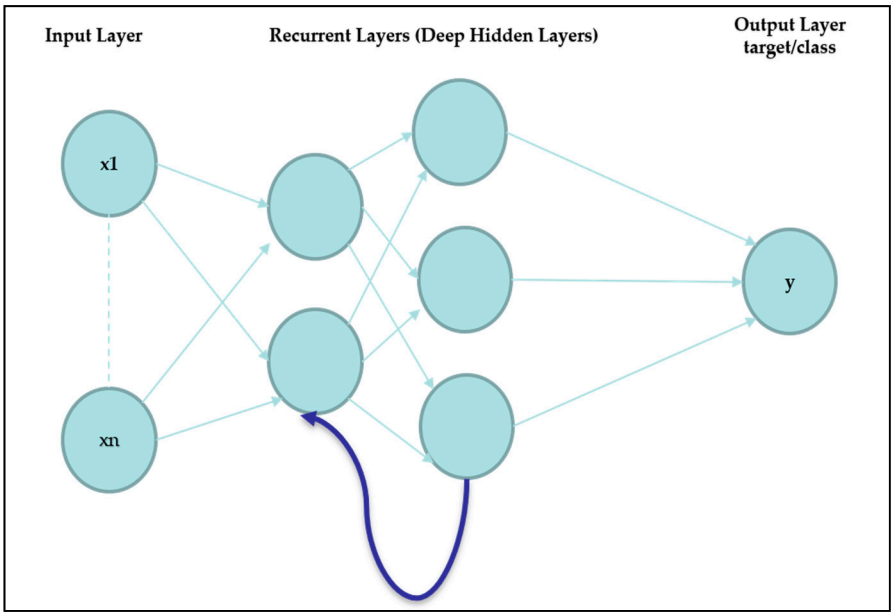

Figure 4. Basic deep neural network architecture.

Simply defining an RNN, we can elaborate it in terms of (∑, S, δ) where ∑ represents the inputs, S represents the states, and δ is the transition function of the neural network. For example, a traditional language model based on RNN considers a document as a sequence. To predict the next word, an LSTM is trained which also takes into account the previous words. Hence, it maximizes p (wt|wt−1|, wt−2, . . . , w0; model). The input words are transformed to vectors in Σ which eventually is used for the LSTM state update. For the output, a projection of s_t is required into a vector of the size of the dictionary and is followed by an activation function, e.g., Softmax. However, challenges occur with more extensive size dictionaries.

3.3. Particle Swarm Optimization (PSO)

The PSO was first proposed by J. Kennedy and R. Eberhart in 1995 [48]. It is a population-based optimization algorithm, which became very famous because of its continuous optimization process towards the best solution.

It is derived from the concepts of swarming habits of animals, e.g., fish or birds and also from genetic algorithms. At a given time *t*, PSO upholds multiple possible solutions, each represented by a particle. The fitness of these solutions is calculated during each iteration by using an optimization function. There is a certain velocity which each particle has to move with, in order to reach the maximum value, returned by the objective function. At each iteration, the particle's position and velocity are updated according to the following Equations (1) and (2):

$$V(t+1) = W * V(t) + c1 * r1 \times [Y(t) - X(t)] + c2 * r2 \times [G(t) - X(t)] \quad (1)$$

$$X(t+1) = X(t) + V(t+1) \quad (2)$$

where V is the particle's velocity, X(t) is particle's current position at time t. Y(t) is the individually best solution of the particle at time t, and G(t) is the global best solution of the swarm at time t. W is the coefficient of inertia, usually ranges between 0.8 to 1.2. r1 and r2 are the random numbers generated in the range [0,1], and c1 and c2 are the cognitive and social coefficients, respectively. These coefficients are also known as learning factors, and their value is usually kept as 2.

4. Hybrid Approach Based on RNNs and Topic Modeling (LSTM-LDA)

In this section, we present a detailed explanation of the proposed approach. Our proposed approach bridges the gap between traditional topic models and LSTM. The primary purpose of this proposed approach is to focus on the strengths of the two models and overcome the challenges faced by them. Therefore, an ideal model for such a task should have quality features, such as less number of parameters, easy to interpret, and able to accurately predict for future trends.

Hence, we can define the joint model as the following Equation (3):

$$\log\ \mathrm{p}(w) = log \sum_{z1:T} \prod_{T}\ \mathrm{p}(w\mathrm{t}\,|z\mathrm{t})\mathrm{p}(z\mathrm{t}\,|z\mathrm{t}-1,\ z\mathrm{t}-2,\ \ldots,\ z1) \tag{3}$$

Model Structure

This section explains the complete structure of the model. Here in the proposed model, LSTM is used to model the topic sequences, i.e., p (zt|zt − 1, zt − 2, . . . , z1), while LDA models word sequences, i.e., p (wi|zi). The parameters for LDA are presented in Table 1.

Table 1. LDA parameters and definitions.

Parameters	Type	Definition
K	Integer	Number of topics
V	Integer	Dictionary/Vocabulary size
D	Integer	Number of documents
N	Integer	Total number of words in document d
alpha	K-dimensional vector [+ve real]	Prior weight of a topic k in a document
beta	V-dimensional vector [+ve real]	Prior weight of a word in a topic
Theta	Float [0,1]	Probability
z	N-dimensional vector of integers	Topic assignments

Let us consider we have topics K, dictionary of size V, collection of documents D where each document d consists of N words. By using all these representations, we can formally define an algorithm for our hybrid model, as shown below in Algorithm 1.

Algorithm 1: Generative model for the hybrid approach

for topic $k = 1 \rightarrow K$
select topic $\varphi_k \sim \mathrm{Dir}(\beta)$
for document $d = 1 \rightarrow D$
    LSTM model initialization with $s_0 = 0$
    for words at $t = 1 \rightarrow N$ in document d
LSTM update as $s_t = \mathrm{LSTM}\ (z_{d,t-1}, s_{t-1})$
Calculate $\theta = \mathrm{softmax}_K(W_p s_t + b_p)$ (for topic distributions)
        Select a topic $z_{d,t}$ from θ
        Select word $w_{d,t}$ from $\varphi_{zd,t}$

Subsequently, we compute the probability of the topic for the next word based on the topics of preceding words. In Figure 5, the architectural overview of our proposed model is presented. The input w(t) is the word vectors generated for the comment at time t, and z(t) is the latent vector generated as a result of the LDA process. This latent vector z(t) is fed to the LSTM model. After training, LSTM calculates the topic category of any given comment. The ground truth data is composed of comments along with category labels.

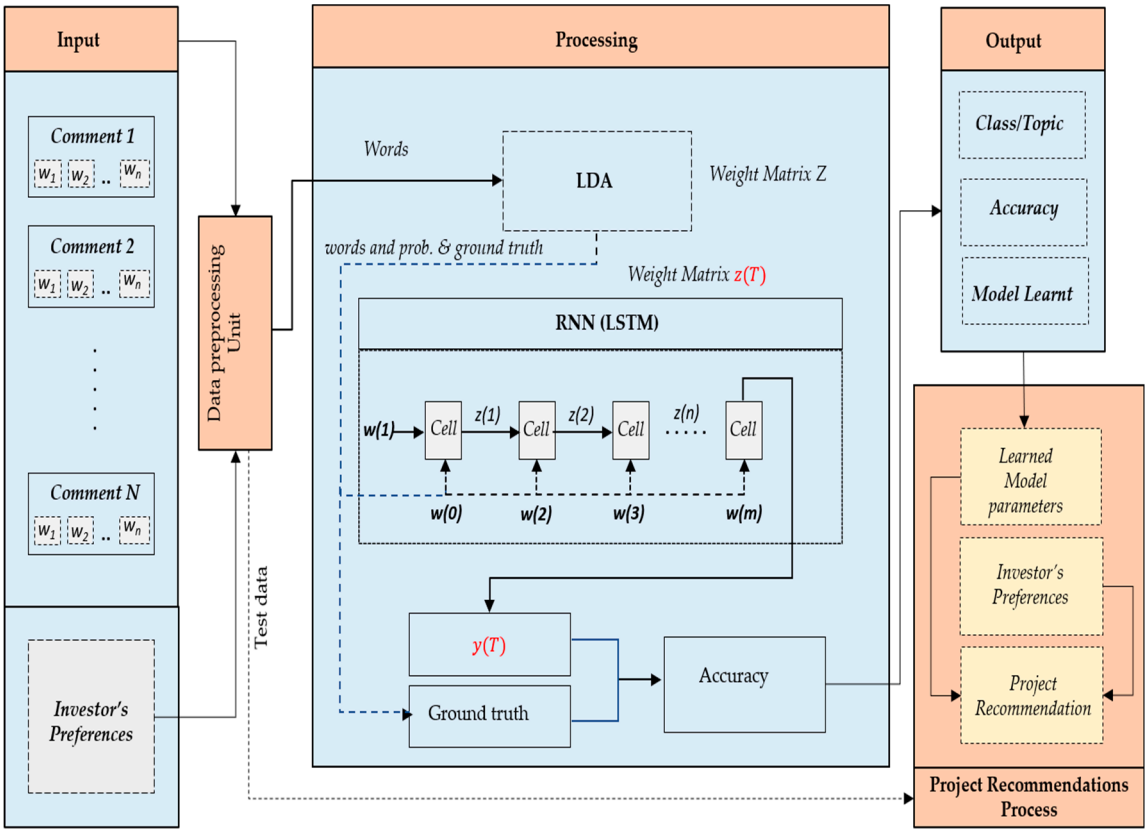

Figure 5. Proposed system architecture for topic class prediction and optimized recommendations.

The prediction module uses the LSTM based topic model to predict the topic class for the next word in the document. The input comments are preprocessed and are converted to word vectors. LDA learns topics, and LSTM is trained based on this pre-trained model. The second module is the recommendation module, which suggests some projects based on the topics discovered by incorporating the investor's preferences.

5. Implementation Environment

In this section, we present the details of the implementation environment. It covers experimental setup details, explanation of data collection process, data description, and model structure.

5.1. Experimental Setup

The experimental setup is summarized in Table 2. The core system components are Ubuntu 18.04.1 LTS as an operating system, with 23Gb memory, GPU is Nvidia GForce 1080. The implementation is done in Python language along with Tensorflow API. For all experiments, we have divided the data into 70% training data and 30% testing data. By training the 70% of the data, we were able to have ground-truth data ready for predicting the topics and trends of discussions in the remaining testing data. The primary objective is to discover the trends in the comments and recommend investors accordingly.

Table 2. Implementation and experimental environment.

System's Components	Specifications
Operating System	Ubuntu 18.04.1 LTS
Memory	32Gb
Language	Python
GPU	Nvidia GForce 1080
Language Version	3.6.1
API	Tensorflow
API Version	1.13

5.2. Data Set

In this section, we present the data collection and selection process.

5.2.1. Data Collection

For the experiments, we targeted the comments section of the Kickstarter campaigns. To collect ground truth data, we need comments for both the categories under analysis, i.e., scam/potentially fraudulent-campaigns and non-scam/genuine campaigns.

There is no public data available for scam cases in crowdfunding, so we hand-collected data from the most substantial reward-based and leading crowdfunding platform—Kickstarter. Kickstarter's policies are comparatively stringent to other crowdfunding platforms. Though the probability of fraudulent activities drops down due to Kickstarter's strict policies, there are a reasonable number of campaigns reported or discovered as suspicious or potential frauds. To come up with reasonably good enough data, we went through extensive research and analysis.

The data collection process can be divided into three primary phases, as shown in Figure 6, one involving extensive research based on news, e.g., CNNmoney.com, articles, discussion forums, e.g., reddit.com; social network sites, e.g., facebook.com, and others. This task helped us to create a list of project URLs reported as potential scams or suspended due to some ambiguities or doubtful material presented by the project creators. Our data covers all the authentic and potential fraudulent campaigns published on Kickscammed.com. However, Kickscammed.com does not necessarily cover all the fraudulent cases, so we complement our dataset with news search using Google, Reddit, Facebook, News, and CNN money.

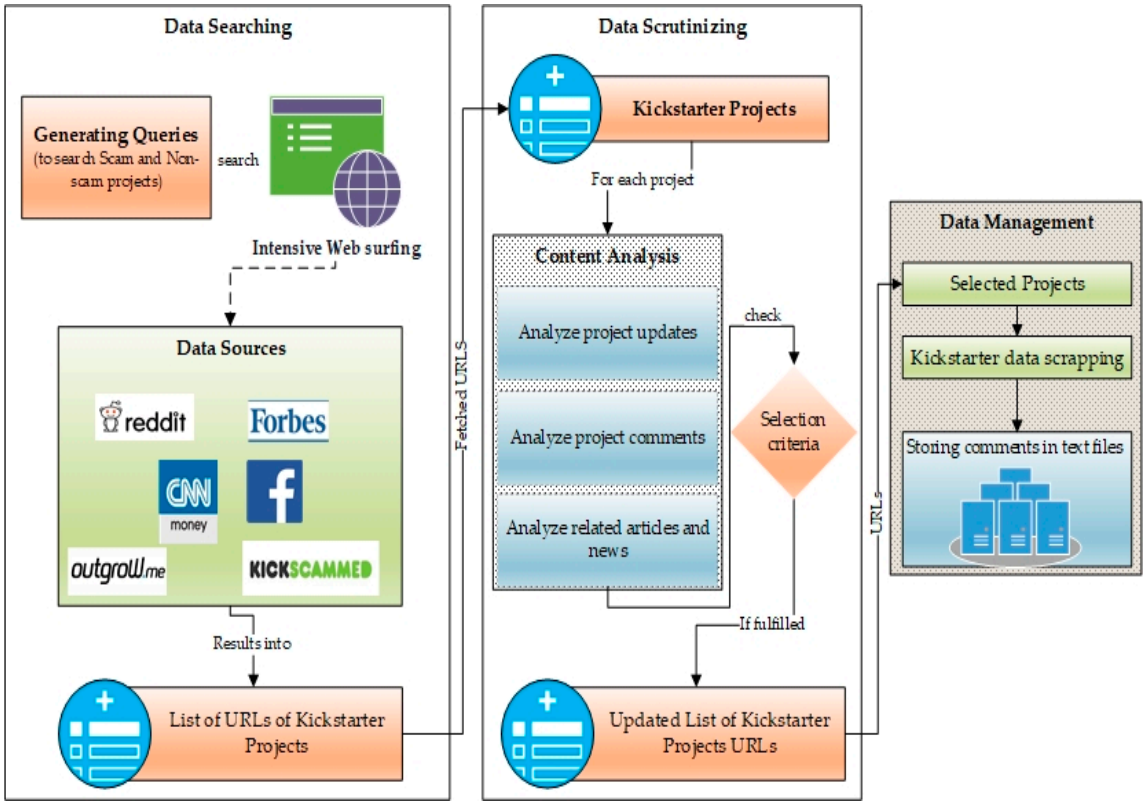

Figure 6. Data collection and selection process.

5.2.2. Experimental Data

In Table 3, the details of the data used in experiments are presented. We have 645,251 comments in total, collected from 600 Kickstarter projects in different categories. The average number of comments per project is 841. We filtered out the comments for which the content was hidden and not available for the public. After comments filtering, total comments were around 504,000 comments. For training, 70% of the entire data was used and the rest was used for testing.

Table 3. Implementation and experimental environment.

Data Characteristics	Specifications
No. of projects	600
Total no. of comments before filtering	645,251
Total no. of comments after filtering	504,184
The average number of comments per project	841
Training data	70%
Test data	30%

5.3. LSTM-LDA Model Training

We accomplish a comprehensive assessment of our model in comparison with different traditional and deep topic models. These experiments are performed on Kickstarter's comments. The training data was used to extract latent topics in comments. After the successful training, 12 topic classes were extracted, which classify the comments appropriately into these topic clusters. As shown in Table 4, the topics are labeled as Topic_0, Topic_1, Topic_2, Topic_3, Topic_4, Topic_5, Topic_6, Topic_7, Topic_8, Topic_9, Topic_10, and Topic_11. After LSTM training, each comment in a project is classified into one of these topic classes.

Table 4. Identified topic classes after LDA analysis.

	Data	Explanation
Topic_0	Waiting for rewards	Fulfill, rewards, waiting
Topic_1	Asking for refunds	Fulfill, refunds, money, creator, project
Topic_2	Waiting for an update/reward	Waiting, money, refund, update, does not receive
Topic_3	Reporting or taking legit actions against it	Attorney, Kickstarter, project, report, actions, response, state, legal
Topic_4	Product never received	Never, still, product, received, mine
Topic_5	Showing anger or disappointment	A fraudster, what, why legally
Topic_6	No communication/confused	Still, no, what
Topic_7	Product shipment	Product, shipped, wedge, idea
Topic_8	Product description	Brewer, cup, drink, work, lid, router, device
Topic_9	Product's working status	Apps, device, excellent, support, ads
Topic_10	Product received	Cards, today, mine, received, loved
Topic_11	Showing excitement	Pledge, received, mine, cards, deck, decks, loving, loved, great

The topic classes from Topic_0 to Topic_6 represent comments related to strongly negative and aggressive emotions of the investors. The rest of the topic classes, i.e., Topic_7 to Topic_11, accommodate neutral or positive sentiments of the investors. Projects falling into the top 7 topic classes (i.e., Topic_0–Topic_6) are considered comparatively less authentic or suspicious, whereas the projects falling into the rest of the categories (i.e., Topic_7–Topic_11) are considered comparatively authentic.

5.4. Optimized Recommendation Module

The goal of the optimized recommendation module is to recommend the best and credible project to the investors by utilizing the results from the topic predictions along with investor's preferences. The process can be described as follows:

1. At step 1, projects with higher authenticity are listed.
2. Then the user's preferences are checked. User preferences include project category, e.g., arts or comics, project location, delivery time, and rewards.
3. At the final step, recommendations of ongoing projects are presented to the user based on their interests.

In order to find projects with the highest authenticity, an optimization function is required. Figure 7 illustrates the overall process flow for optimized project recommendation, in which the recommended projects, along with user preferences, are fed to the optimization module. We are using PSO as an optimization algorithm. It takes user preferences, predicted topic classes, recommended projects, and set of constraints as input.

An objective function is vital for an optimization algorithm to present the optimal output. In our scenario, the objective function will try to find the most authentic project to recommend according to user preferences.

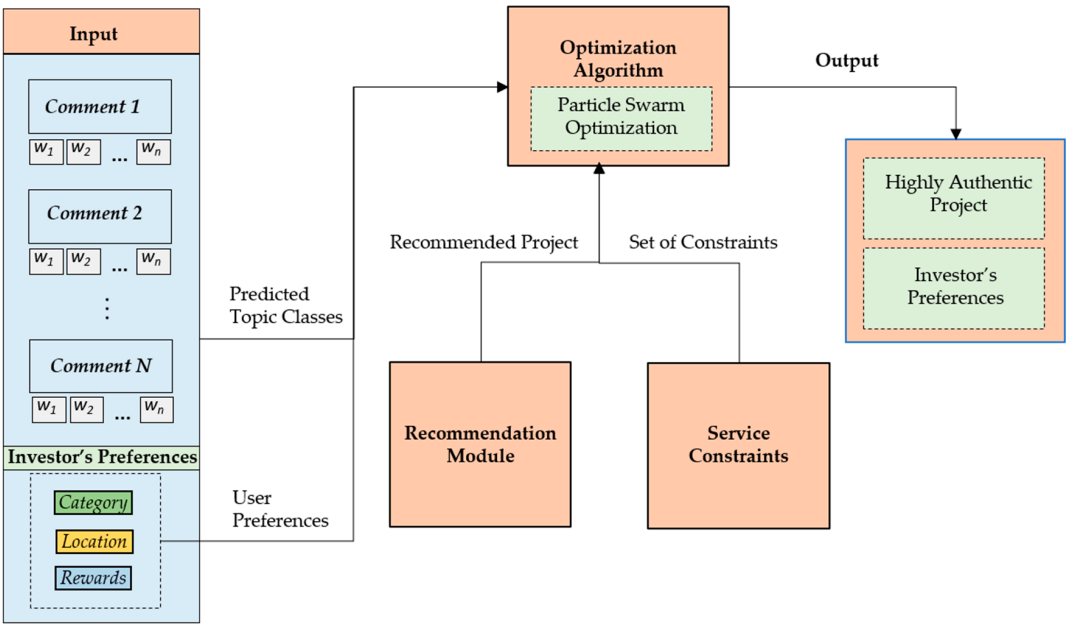

Figure 7. Optimized project recommendation module based on Particle Swarm Optimization (PSO).

5.4.1. Project's Credibility Estimation

In this scenario, we want an objective function to find out the most credible and optimal project recommendation. The most credible project can be defined as a project which falls under user-preferred categories with the highest probability of on-time delivery. We describe and associate a project's credibility with the level of authenticity it has. The authenticity of a project is derived from the patterns of communication in the context of project creator's updates related to the development of the project, keywords used, delivery or reward promises, investors' sentiments, etc. The authenticity of a project is calculated according to Equation (4). We have used the profile related features of the project creator and the content associated elements of a project.

$$Authenticity_{\ project} = \left[\sum_{i=1}^{n} \frac{Class_{ci}}{Class_{Bi}} + \left(\frac{links_{social}}{R_{score} + delay_{post}} \right) \right] - (X_w + Y_w) \tag{4}$$

where, X_w and Y_w are the weights associated with class A, and presence of profile pictures, respectively. The definition of each parameter is given in Table 5.

Table 5. Definitions of the parameters of authenticity.

Parameters for Authenticity		Explanation	Notations
		Content-based	
1	Class A	Weightage of comments in Topic_3 to Topic_5	X_w
2	Class B	%age of comments in Topic_0 to Topic_2 & Topic_6	$Class_B$
3	Class C	%age of comments in Topic_7 to Topic_11	$Class_C$
4	Readability score	The measure of content clarity	R_{score}
		Profile-based	
5	Profile picture	If the project creator has a display picture or not	Y_w
6	Number of social links	External links on profile, e.g., facebook	$links_{social}$
7	Delay between posts	Avg. delay between updates/comments	$delay_{post}$

A project can have multiple comments falling into different topic classes. Therefore, we find the percentage of comments in each topic class. For simplicity, we have divided all the comments into

three primary types, named as class A, class B, and class C. For negative comments, we have two representative classes as class A and class B while all the other comments are represented by class C. This classification is performed based on the criticality and influence of negative comments towards project authenticity.

The nature of the comments in class A is very threatening and unfavorable. Therefore, we have treated this class separately to curtail the risks of unsafe recommendations. Thus, if the percentage of comments in class A is more significant than zero, the value of X_w becomes 1, else it remains zero. Our objective function tries to maximize the percentage of class C and searches for the maximum number of social links provided by the project creator.

For a project to be successful, creator information is as crucial as the content because it increases the trust of an investor. Therefore, we can infer that the presence of profile picture and external links provided by a project creator are directly proportional to the authenticity of a project. Thus, the presence of a profile picture means, the value of Y_w is 0; else its value is 0.5. One common feature of the project creator is related to his patterns of updates and comments referred to as $delay_{post}$. It represents the average gap or delay between any two consecutive updates or comments posted by the creator. It shows the involvement or communication rate of a creator towards the project development. Therefore, an increase in $delay_{post}$, will harm a project's authenticity. The values from the above equation are normalized between 0 to 1, where 0 means highly unauthentic, and 1 means highly authentic. It depicts how much trust one can put into a project. Therefore, the higher the authenticity is, the more credible a project becomes as described in the Equation (5).

$$Authenticity_{\ project} \propto \ Credibility_{project} \tag{5}$$

Consequently, we have five different levels of credibility as extremely low, low, normal, high, and extremely high, falling into varying degrees of authenticity range, as shown in the Table 6.

Table 6. Credibility (in terms of authenticity levels) of projects with example scenarios.

Project's Credibility	Example Scenarios	Representative Topic Classes	Authenticity Range
Extremely Low	• Investors are heading towards taking legal action against the project creator. • If no one has received the product or reward for more than a year. • Investors are furious towards project creators.	Topic_3, Topic_4, Topic_5	[> 0 ≤ 0.3]
Low	• Promised rewards are still pending • Investors did not get refunds • Lack of communication as no updates or comments left by the creator after successful funding	Topic_0, Topic_1, Topic_2, Topic_6	[> 0.3 ≤ 0.6]
Normal	• No harsh sentiments exist by the investors • Positively waiting for the product • Some have received updates or rewards	Topic_8	[> 0.6 ≤ 0.7]
High	• Investors are happy and excited about the product • Hopes are very high • Receiving updates	Topic_7, Topic_11	[> 0.8 ≤ 0.9]
Extremely High	• Investors have received the product • Talking about the working conditions of the product	Topic_9, Topic_10	[>0.9 ≤ 1.0]

The projects in shallow and low credibility class will have higher risks of being forged. In other words, that project has maximum chances of non-payments, non-delivery, lack of communications, and lack of responses in terms of comments or updates from the project creator. Hence, investing in such a project is not recommended. On the other hand, a project with high or extremely high credibility is an eminently favorable project to invest in, as it is more likely to get delivered on time.

5.4.2. Objective Function for Optimal Project Recommendations

In our case, the goal of the objective function is to find a project with a maximum user preference value and maximum authenticity, i.e., a project with high or extremely high credibility.

Therefore, we need to design an objective function which fulfills the following criteria:

1. Maximizes the higher credibility levels (i.e., maximizes weights for Topic_7, Topic_9, Topic_10, and Topic_11)
2. Maximizes user preferences.
3. Minimizes the lower credibility level (i.e., minimizes weights for Topic_0, Topic_1, Topic_2, Topic_3, Topic_4, Topic_5, and Topic_6).

Thus,

$$weight_1 = \alpha(Topic\_7) + \beta\,(Topic\_9) + \delta\,(Topic\_10) + \Upsilon\,(Topic\_11) + \omega\,(\text{User preferences}) \quad (6)$$

$$weight_2 = \Phi(Topic\_0) + \Phi 1(Topic\_1) + \Phi 2\,(Topic\_2) + \Phi 3\,(Ttopic\_3) + \Phi 4\,(Topic\_4) + \Phi 5\,(Topic\_5) + \Phi 6\,(Topic\_6) \quad (7)$$

In Equation (6), α, β, δ, Y and ω are the weights associated with *Topic*_7 (product shipment), *Topic*_9 (product working status), *Topic*_10 (product received), *Topic*_11 (showing excitement), and user preferences respectively. Similarly, in Equation (7), Φ, $\Phi 1$, $\Phi 2$, $\Phi 3$, $\Phi 4$, $\Phi 5$, and $\Phi 6$, are weights associated with *Topic*_0 (waiting for rewards), *Topic*_1 (asking for refunds), Topic_2 (waiting for an update/reward), *Topic*_3 (reporting or taking legit actions), Topic_4 (product never received), *Topic*_5 (showing anger or disappointment), and *Topic*_6 (no communication/confused), respectively. The target of the objective function in PSO is to minimize the weights of topic classes from 0 to 6; and maximize the weights of user preferences along with topic classes 7, 9, 10, and 11. Hence, the objective function can be described as Equation (8) below:

$$weight = Max(weight_1) + Min\,(weight_2) \quad (8)$$

Figure 8 shows the working flow of the PSO algorithm. In PSO, particle velocities are randomly initialized for the generated population. In the next step, their positions are randomly assigned; and based on these current positions and velocities, the fitness of each particle is estimated. Thus, it can help us formulate the objective function for the fitness estimation as it considers both the maximization function and the minimization function. In the next step, the current fitness value of the particle is compared with its individual best fitness, i.e., Y(t). If the current fitness value is better than Y (t), Y(t) is updated with current fitness value else it remains the same. After this, Y (t) is compared with the global fitness, i.e., G(t). If G(t) is better than Y(t), Y(t) is updated to G(t), else it remains the same. For each particle, its velocity and positions are updated in each iteration to the best fitness values.

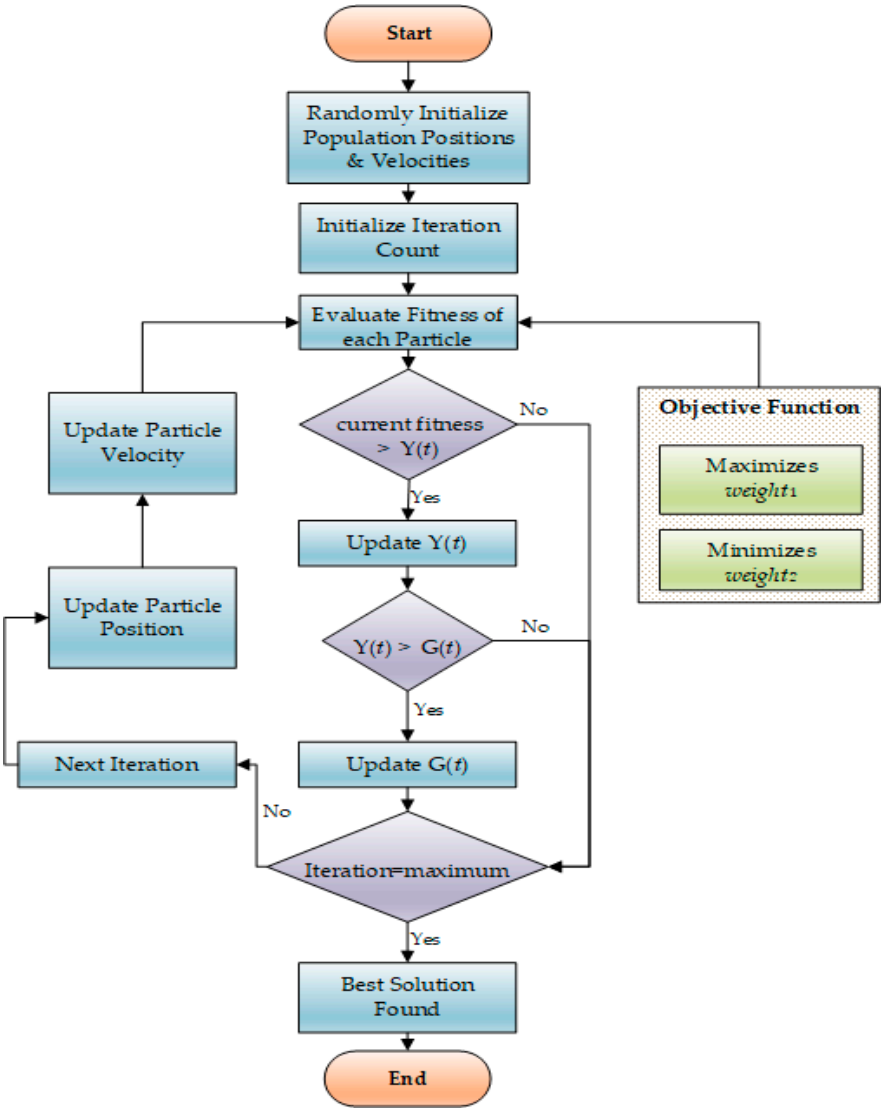

Figure 8. Flow chart for optimized recommendations based on PSO.

6. Results

In this section, we present the analysis of the results of our hybrid approach towards topic class prediction in crowdfunding comments. Based on these predictions, we implemented an optimized recommendation mechanism that recommends the highly authentic project to users under their preferences. In Section 6.1, we present the initial prediction accuracy of LSTM-LDA.

6.1. Optimized Recommendation Module Prediction Accuracy

This section shows the essential topic class prediction accuracy of our model as compared with other baseline algorithms named as simple Neural Networks (NNs), NN-LDA.

6.1.1. Topic Class Prediction Accuracy

In Figure 9, the graph presents the results of the prediction accuracy of our proposed approach with other baseline approaches. We have compared LSTM-LDA (RNN-LDA) with a basic NN model and LDA based NN model. We can observe that for LSTM-LDA, the prediction accuracies are better, i.e., approximately 96% as compared with NN (approximately 92%) and NN-LDA (about 95%).

Though NN-LDA and LSTM-LDA accuracy are pretty close, LSTM-LDA has a more stable pattern as compared with NN-LDA. The prediction accuracy for RNN-LDA gets steady after 300 epocs.

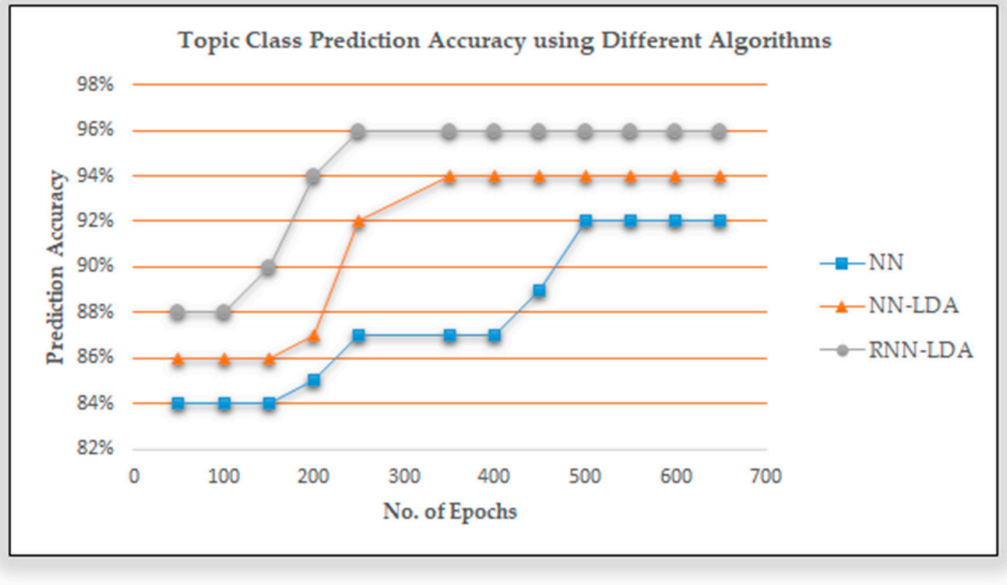

Figure 9. Topic class prediction accuracy for different algorithms.

6.1.2. Topic Class Prediction Accuracy for a Varying Number of Topics

Here, we experimented with the different number of topics for LDA training, for topic class prediction of text data. This way, we found the optimal number of topics for better predictions as well. We varied the number of topics between 5 and 20.

As shown in Figure 10, the prediction accuracies change with a different number of topics. For all the algorithms used, we can observe that maximum accuracy is achieved when the number of topics is between 10 and 15. If the number of topics is 12, all the algorithms achieve their highest prediction accuracies.

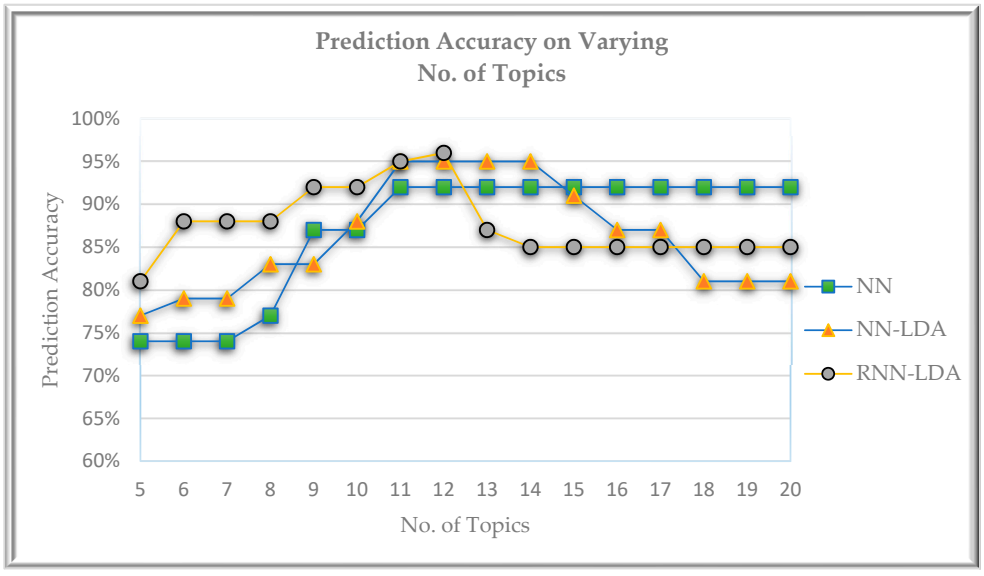

Figure 10. Prediction accuracy for a varying number of topics.

6.1.3. Discussion Trends in Suspicious Campaigns over Time

In this section, the experiment is performed to discover discussion trends in crowdfunding campaigns. For this experiment, we selected suspicious campaigns. Suspicious campaigns are characterized as campaigns with low or extremely low credibility levels.

To verify the trends over time, we chose different phases of a campaign. The campaign's lifetime was divided into four phases, i.e., during the funding period phase, between the funding period and expected delivery phase, after one month of the expected delivery phase, and after one or more than one year of the expected delivery phase.

Figure 11 shows that during the funding period, the motivation level of backers is quite high. They post good stuff revealing their emotional level; after the funding period passed, the excitement level drops from 45 to 27%. The reason we observed is when the project enters into the implementation phase, the creator's involvement starts decreasing, i.e., lack of updates or comments start raising other emotions than excitement in backers. Once the delivery date has passed, and there is no communication from the creator's end, the excitement turns into disappointment and anger. The content of the comments is outstandingly consisting of accusations, allegations, frustrations about not receiving the product or any update from the project creator, anger, refund or reward claims, etc.

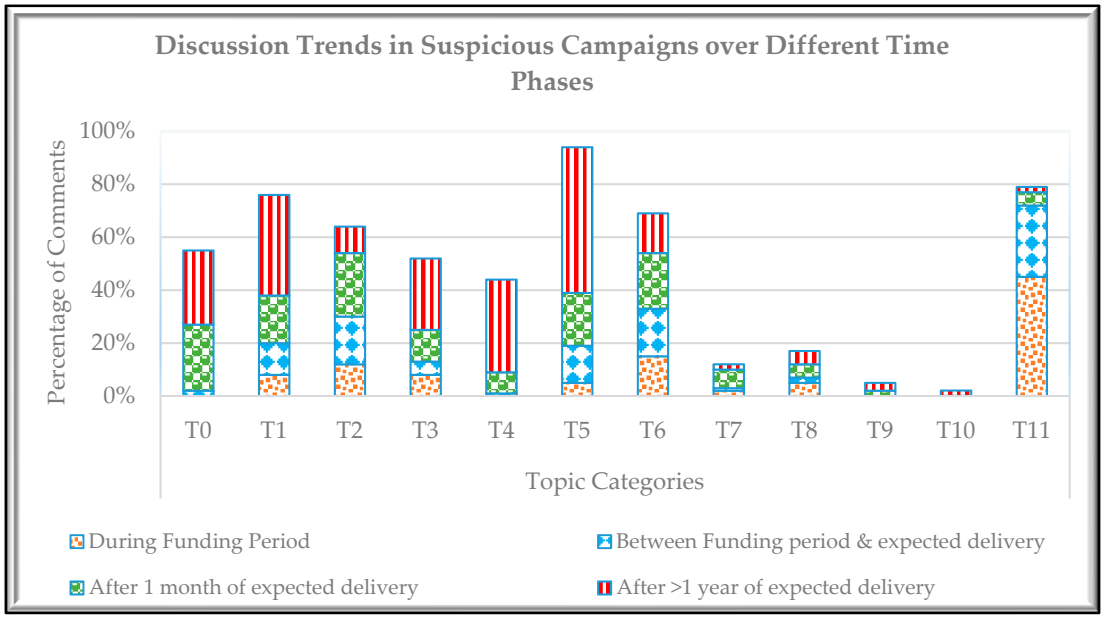

Figure 11. Discussion trends over time in suspicious campaigns based on topic classes.

7. Discussions and Concluding Remarks

This section emphasizes the challenges, implications, and limitations in the development of the proposed recommendation system in crowdfunding field. In this study, an optimized recommendation system is developed to help investors in the selection process of authentic and reliable projects by accommodating their interests. We have presented a hybrid model of RNN-LSTLM and topic modeling that combine the benefits of both LSTMs, which can capture time dependencies for the topic and class prediction, and topic modeling, which can extract topics that can help predict scams. Our proposed model has improved the accuracy of scam prediction significantly from the baseline LDA-based models. We have also embedded an optimized recommendation strategy based on a project's credibility.

There were several challenges in the development of the proposed system. One of the key challenges was collection of ground truth data, and its verification, was a tedious task. As there is no public data available for scam cases in crowdfunding, so we hand-collected data from the most

substantial reward-based and leading crowdfunding platform—Kickstarter. Additionally, all the relevant data and clues were cross-checked manually several times, which was quite a time taking task.

The primary goal of our recommendation systems is to determine the credibility of a project before recommending it to the investors by incorporating their preferences. Online markets, e-commerce, and all form of online business exchanges involve risks and uncertainty. Similarly, for crowdfunding platforms, it is very crucial to establish or build investors' trust to curtail the chances of uncertainty and risky investments. Previous studies show that many elements are vital for trust-building in any domain, such as content quality and readability [49] and presence of profile picture [50]. Therefore, to determine the credibility of a project, we evaluate essential elements such as the profile of the creator, his or her communication patterns and sentiments of investors. The projects with highest credibility level are safe to recommend. However, in case of crowdfunding projects, the amount of risk an investor can tolerate becomes subjective. If an investor is not concerned about the money and only interested in the idea, he or she can invest money to bring that idea to life. Therefore, we present the authenticity scores of projects and evaluate their credibility accordingly. The rest is in the investor's hand to make a decision accordingly. Some theoretical and practical contributions of the study can be summarized as follows.

First, our results complement the previous studies [42,43] and show that the hybrid approaches perform better in topic identification. With LSTM-LDA, we were able to achieve 96% accuracy in topic predictions. We also evaluated these topic classes to help identify suspicious campaigns.

Second, we investigated to discover how configuring comments along with project timelines can help improve the prediction quality. For this experiment, we divided the comments into five different batches described in Table 7. Each batch represents a specific timeline, such as batch 1 represents the period between the campaign launched until fundraising. Similarly, other batches are created based on different development phases of a project. There are 50–70 comments in each batch. Projects with less than 50 comments were dropped out for this experiment. From Table 7, we can observe that the most recent comments in a project are significant towards an accurate classification of scam or non-scams. Additionally, the comments discovered in the latest batch, i.e., batch 5, are pure towards one topic class. In other words, they are less likely to have a mixture of discussions as people are generally sharing the same thoughts in recent comments.

Table 7. Project classification accuracy based on different comment batch sizes.

Batch No.	Period	Accuracy
Batch 1	Launched until Fundraising period	20%
Batch 2	After Fundraising period	55%
Batch 3	Before expected delivery	67%
Batch 4	After Expected Delivery	83%
Batch 5	Top 60 most recent comments	88%

Third, there is plenty of data on the internet, which is beneficial for almost every organization, beating informational uncertainty and economic risks remains a hefty challenge for people. There is an ample amount of work done on user-generated content, such as comments and reviews, for different domain-specific applications. However, in crowdfunding, these comments and reviews are often ignored. Due to an increasing threat of fraudulent and suspicious actions on such forums, accurate and timely analysis of the content generated by the stakeholders can help us mitigate these risks. Therefore, we are finding a project's credibility based on both the content and behavioral patterns of the creator.

Fourth, the optimized recommendations based on the project's credibility is also a significant addition in the crowdfunding domain. Moreover, this module gives weight to the user preferences and helps them invest in the most credible project.

In Table 8, a comparison of recent works in crowdfunding with our proposed approach is presented. Here, we have compared our approach and interests with few past recent papers (since 2015) in

crowdfunding. In [51] and [52], different crowdfunding domains are targeted, such as tourism and medical crowdfunding, respectively. Though both of these studies have attempted risk assessments and fraud identification, none of these studies focus on the project content or linguistic features. In [53–57], Kickstarter is used as a crowdfunding platform primarily for project success predictions. The work in [58,59] and [26], targets to acquire knowledge of fraudulent behaviors in crowdfunding by analyzing linguistic patterns. In [60], comments and updates of successful projects are used to estimate the delivery duration of rewards. In comparison with all the recent works, and to the best of our knowledge, the proposed approach is first of its kind in considering investors' comments and utilizing them for optimized and reliable recommendations.

Table 8. Comparison analysis of some recent studies in crowdfunding with our proposed approach.

Works	Platform	Language Analysis	LDA	LDA-LSTM	Optimization	Comments	Fraud Detection
[26]	Kickstarter	√	X	X	X	X	√
[51]	Tourism Crowdfunding	X	X	X	X	X	X
[52]	Medical Crowdfunding	X	X	X	X	X	√
[53]	Kickstarter	X	X	X	X	X	X
[54]	Kickstarter	√	X	X	X	X	X
[55]	Kickstarter	√	√	X	X	X	X
[56]	Kickstarter	√	√	X	X	X	X
[57]	Kickstarter	√	√	X	X	X	X
[58]	Kickstarter	√	X	X	X	X	√
[59]	Kickstarter	√	X	X	X	Count only	√
[60]	Kickstarter	X	X	X	X	√	X
Proposed approach	Kickstarter	√	√	√	√	√	√

The idea of using user reviews and comments for prediction and recommendation is not novel. Many studies have been conducted, which make use of reviews to tackle different problems in diverse fields. Some studies have focused on the reviews content to find the satisfaction of hotel guests [61], tourist satisfaction [62], and sentiments related to a product [63]. Therefore, we compared our approach with other recent approaches in Table 9. All of the studies [61–68] are very recent and are targeting different problem domains by using online reviews. All the studies either detect sentiments or the credibility of a review. Our proposed approach is based on a deep contextual understanding of the comments for discovering discussion patterns and making recommendations accordingly.

However, there are some limitations to our work that can be overcome in future research.

First, we have not taken all the profile related features of project creators into consideration. For example, the history of their work on Kickstarter in terms of the number of projects they have been part of (either as a creator or an investor) can tell a lot about their credibility.

Second, it is very challenging to evaluate the risk for each user. There might be some investors who have a great interest in a specific project, and they are ready to take any monetary risk to try that project regardless of the credibility of a project. In that case, a project with an adverse credibility level will less likely be suggested.

Third, we have used data only from one crowdfunding site Kickstarer.com, which is a leading reward-based platform. Hence, it requires to be verified on other platforms to check if the results can be generalized or not. Additionally, considering other platforms can increase dataset size, which can improve the efficiency of our recommendation system.

Fourth, we considered a limited number of features or variables. For example, for credibility assessment, we considered readability score. However, in the future, we can use other linguistic features such as use of pronouns, adjectives, and expressiveness. We aim to address these limitations and challenges in the future.

Table 9. Comparison analysis of some recent studies on reviews with our proposed approach.

Works	Data Source	Scam Detection	Methodology	Product	Optimization	Outcome
[64]	Doctor search portal	√	Deep learning	Doctors	X	Review classification
[61]	TripAdvisor, Expedia, Yelp	X	Linguistic analysis	Hotels	X	X
[65]	TripAdvisor	X	Cloud model for probabilistic linguistic information	Hotels	X	Review Summarization
[66]	Yelp	X	Linguistic Style Matching (LSM)	Restaurants	X	Emotional dimensions
[67]	Cornell sentiment polarity dataset	√	Sentiment analysis and machine learning	Movie reviews	X	Review classification
[68]	Customer reviews	√	Game theory	Shopping	X	Review classification
Proposed approach	Kickstarter	√	Topic Modeling, LSTM, PSO	Secure Crowdfunding projects	√	Project classification based on comments

Author Contributions: W.S. conceived the idea for this paper, designed the experiments, wrote the paper, assisted in algorithms implementation; design and simulation; Y.-C.B. conceived the overall idea of this paper, proof-read the manuscript and supervised the work.

Funding: This research was financially supported by the Ministry of SMEs and Startups (MSS), Korea, under the "Regional Specialized Industry Development Program(R&D or non-R&D, Project Number: P0003167)" supervised by the Korea Institute for Advancement of Technology (KIAT); and by 'Jeju Industry-University Convergence Foundation' funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea). [Project Name: 'Jeju Industry-University convergence Foundation/Project Number: N0002327].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World's Internet Users Report 2018. Available online: https://wearesocial.com/blog/2018/01/global-digital-report-2018 (accessed on 10 October 2019).
2. Fang, B.; Ye, Q.; Kucukusta, D.; Law, R. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tour. Manag.* **2007**, *52*, 498–506. [CrossRef]
3. Internet Crime Report 2018. Available online: https://www.fbi.gov/news/stories/ic3-releases-2018-internet-crime-report-042219 (accessed on 10 October 2019).
4. 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, Seattle, WA, USA, 11–15 February 2012; p. 10.
5. Communicating Science Online Increases Interest, Engagement and Access to Funds. 2019. Available online: https://theconversation.com/communicating-science-online-increases-interest-engagement-and-access-to-funds-122102 (accessed on 10 October 2019).
6. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
7. Peng, D.; Guilan, D.; Yong, Z. Contextual-LDA: A Context Coherent Latent Topic Model for Mining Large Corpora. In Proceedings of the IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, Taiwan, 20–22 April 2016; pp. 420–425.
8. Block, J.; Hornuf, L.; Moritz, A. Which updates during an equity crowdfunding campaign increase crowd participation? *Small Bus. Econ.* **2018**, *50*, 3–27. [CrossRef]
9. Mollick, E. The dynamics of crowdfunding: An exploratory study. *J. Bus. Ventur.* **2014**, *29*, 1–16. [CrossRef]
10. Kuppuswamy, V.; Bayus, B.L. Crowdfunding creative ideas: The dynamics of project backers in Kickstarter. *SSRN* **2014**. [CrossRef]
11. Crosetto, P.; Regner, T. It's never too late: Funding dynamics and self pledges in reward-based crowdfunding. *Res. Policy* **2018**, *47*, 1463–1477. [CrossRef]
12. Kuppuswamy, V.; Bayus, B.L. Crowdfunding creative ideas: The dynamics of project backers. In *The Economics of Crowdfunding*; Palgrave Macmillan: London, UK, 2018; pp. 151–182.

13. Marom, D.; Robb, A.; Sade, O. Gender dynamics in crowdfunding (Kickstarter): Evidence on entrepreneurs, investors, deals and taste-based discrimination. *SSRN* **2016**. [CrossRef]
14. Van de Rijt, A.; Kang, S.M.; Restivo, M.; Patil, A. Field experiments of success-breeds-success dynamics. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6934–6939. [CrossRef]
15. Block, J.H.; Colombo, M.G.; Cumming, D.J.; Vismara, S. New players in entrepreneurial finance and why they are there. *Small Bus. Econ.* **2018**, *50*, 239–250. [CrossRef]
16. Povel, P.; Sertsios, G.; Kosova, R.; Kumar, P. Boom and gloom. *J. Financ.* **2016**, *71*, 2287–2332. [CrossRef]
17. Agrawal, A.; Catalini, C.; Goldfarb, A. Crowdfunding: Geography, social networks, and the timing of investment decisions. *J. Econ. Manag. Strategy* **2015**, *24*, 253–274. [CrossRef]
18. Belleflamme, P.; Lambert, T.; Schwienbacher, A. Individual crowdfunding practices. *Ventur. Capital* **2013**, *15*, 313–333. [CrossRef]
19. Ahlers, G.K.; Cumming, D.; Günther, C.; Schweizer, D. Signaling in equity crowdfunding. *Entrep. Theory Pract.* **2015**, *39*, 955–980. [CrossRef]
20. Vismara, S. Equity retention and social network theory in equity crowdfunding. *Small Bus. Econ.* **2016**, *46*, 579–590. [CrossRef]
21. Li, Y.; Rakesh, V.; Reddy, C.K. Project success prediction in crowdfunding environments. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 247–256.
22. Greenberg, M.D.; Pardo, B.; Hariharan, K.; Gerber, E. Crowdfunding support tools: Predicting success & failure. In Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 1815–1820.
23. Chung, J.; Lee, K. A long-term study of a crowdfunding platform: Predicting project success and fundraising amount. In Proceedings of the 26th ACM Conference on Hypertext & Social Media, Guzelyurt, Northern Cyprus, 1–4 September 2015; pp. 211–220.
24. Cheng, C.; Tan, F.; Hou, X.; Wei, Z. Success prediction on crowdfunding with multimodal deep learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; AAAI Press: Palo Alto, CA, USA, 2019; pp. 2158–2164.
25. Mitra, T.; Gilbert, E. The language that gets people to give: Phrases that predict success on kickstarter. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, MD, USA, 15–19 February 2014; pp. 49–61.
26. Shafqat, W.; Lee, S.; Malik, S.; Kim, H.C. The language of deceivers: Linguistic features of crowdfunding scams. In Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; pp. 99–100.
27. Parhankangas, A.; Renko, M. Linguistic style and crowdfunding success among social and commercial entrepreneurs. *J. Bus. Ventur.* **2017**, *32*, 215–236. [CrossRef]
28. Yuan, H.; Lau, R.Y.; Xu, W. The determinants of crowdfunding success: A semantic text analytics approach. *Decis. Support Syst.* **2016**, *91*, 67–76. [CrossRef]
29. Xu, A.; Yang, X.; Rao, H.; Fu, W.T.; Huang, S.W.; Bailey, B.P. Show me the money: An analysis of project updates during crowdfunding campaigns. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May 2014; pp. 591–600.
30. Blei, D.M.; Lafferty, J.D. A correlated topic model of science. *Ann. Appl. Stat.* **2007**, *1*, 17–35. [CrossRef]
31. Hall, D.; Jurafsky, D.; Manning, C.D. Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 363–371.
32. Yasunaga, M.; Lafferty, J. TopicEq: A joint topic and mathematical equation model for scientific texts. *arXiv* **2019**, arXiv:1902.06034. [CrossRef]
33. Ding, R.; Nallapati, R.; Xiang, B. Coherence-Aware Neural Topic Modeling. *arXiv* **2018**, arXiv:1809.02687.
34. Larochelle, H.; Lauly, S. A neural autoregressive topic model. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2708–2716.
35. Miao, Y.; Yu, L.; Blunsom, P. Neural variational inference for text processing. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1727–1736.

36. Miao, Y.; Grefenstette, E.; Blunsom, P. Discovering discrete latent topics with neural variational inference. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 2410–2419.
37. Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. Recurrent neural network-based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 26–30 September 2010.
38. Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the limits of language modeling. *arXiv* **2016**, arXiv:1602.02410.
39. Messina, R.; Louradour, J. Segmentation-free handwritten Chinese text recognition with LSTM-RNN. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 171–175.
40. Wang, Z.; Liu, J.C. Translating Mathematical Formula Images to LaTeX Sequences Using Deep Neural Networks with Sequence-level Training. *arXiv* **2019**, arXiv:1908.11415.
41. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
42. Mikolov, T.; Zweig, G. Context dependent recurrent neural network language model. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012; pp. 234–239.
43. Zaheer, M.; Ahmed, A.; Smola, A.J. Latent LSTM allocation joint clustering and non-linear dynamic modeling of sequential data. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3967–3976.
44. Chen, Z.; Teng, S.; Zhang, W.; Tang, H.; Zhang, Z.; He, J.; Fang, X.; Fei, L. LSTM Sentiment Polarity Analysis Based on LDA Clustering. In Proceedings of the CCF Conference on Computer Supported Cooperative Work and Social Computing, Kunming, China, 16–18 August 2018; Springer: Singapore, 2018; pp. 342–355.
45. Tian, F.; Gao, B.; He, D.; Liu, T.Y. Sentence level recurrent topic model: letting topics speak for themselves. *arXiv* **2016**, arXiv:1604.02038.
46. Serban, I.V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; Bengio, Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
47. Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A.C.; Bengio, Y. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 2015; pp. 2980–2988.
48. Eberhart, R.; Kennedy, J. Particle swarm optimization. In Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
49. Ermakova, T.; Baumann, A.; Fabian, B.; Krasnova, H. Privacy Policies and Users' Trust: Does Readability Matter? In Proceedings of the Twentieth Americas Conference on Information Systems (AMCIS 2014), Savannah, GA, USA, 7–9 August 2014.
50. Xu, Q. Should I trust him? The effects of reviewer profile characteristics on eWOM credibility. *Comput. Hum. Behav.* **2014**, *33*, 136–144. [CrossRef]
51. Kim, M.J.; Bonn, M.; Lee, C.K. The effects of motivation, deterrents, trust, and risk on tourism crowdfunding behavior. *Asia Pac. J. Tour. Res.* **2019**, *25*, 244–260. [CrossRef]
52. Zenone, M.; Snyder, J. Fraud in medical crowdfunding: A typology of publicized cases and policy recommendations. *Policy Internet* **2019**, *11*, 215–234. [CrossRef]
53. Hu, W.; Yang, R. Predicting the success of Kickstarter projects in the US at launch time. In Proceedings of the SAI Intelligent Systems Conference, London, UK, 5–6 September 2019; pp. 497–506.
54. Desai, N.; Gupta, R.; Truong, K. *Plead or Pitch? The Role of Language in Kickstarter Project Success*; Stanford University: Stanford, CA, USA, 2015.
55. Sawhney, K.; Tran, C.; Tuason, R. *Using Language to Predict Kickstarter Success*; Stanford University: Stanford, CA, USA, 2016.
56. Westerlund, M.; Singh, I.; Rajahonka, M.; Leminen, S. Can short-text project summaries predict funding success on crowdfunding platforms? In *ISPIM Conference Proceedings*; The International Society for Professional Innovation Management (ISPIM): Manchester, UK, 2019; pp. 1–15.

57. Do Carmo, R.A.; Kang, S.M.; Silva, R. Visualization of topic-sentiment dynamics in crowdfunding projects. In Proceedings of the International Symposium on Intelligent Data Analysis, London, UK, 26–28 October 2017; pp. 40–51.
58. Siering, M.; Koch, J.A.; Deokar, A.V. Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *J. Manag. Inf. Syst.* **2016**, *33*, 421–455. [CrossRef]
59. Cumming, D.J.; Hornuf, L.; Karami, M.; Schweizer, D. Disentangling crowdfunding from fraudfunding. *SSRN* **2016**. [CrossRef]
60. Tran, T.; Lee, K.; Vo, N.; Choi, H. Identifying on-time reward delivery projects with estimating delivery duration on kickstarter. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney, Australia, 31 July–3 August 2017; ACM: New York, NY, USA, 2017; pp. 250–257.
61. Crotts, J.C.; Mason, P.R.; Davis, B. Measuring guest satisfaction and competitive position in the hospitality and tourism industry an application of stance-shift analysis to travel blog narratives. *J. Travel Res.* **2009**, *48*, 139–151. [CrossRef]
62. Xiang, Z.; Du, Q.; Ma, Y.; Fan, W. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tour. Manag.* **2017**, *58*, 51–65. [CrossRef]
63. Ali, N.M.; El Hamid, A.; Mostafa, M.; Youssif, A. Sentiment analysis for movies reviews dataset using deep learning models. *Int. J. Data Min. Knowl. Manag. Process* **2019**, *9*, 1–9.
64. Shukla, A.; Wang, W.; Gao, G.G.; Agarwal, R. Catch me if you can: Detecting fraudulent online reviews of doctors using deep learning. *SSRN* **2019**. [CrossRef]
65. Peng, H.G.; Zhang, H.Y.; Wang, J.Q. Cloud decision support model for selecting hotels on TripAdvisor.com with probabilistic linguistic information. *Int. J. Hosp. Manag.* **2018**, *68*, 124–138. [CrossRef]
66. Wang, X.; Tang, L.R.; Kim, E. More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? *Int. J. Hosp. Manag.* **2019**, *77*, 438–447. [CrossRef]
67. Elmurngi, E.; Gherbi, A. Detecting fake reviews through sentiment analysis using machine learning techniques. In Proceedings of the Sixth International Conference on Data Analytics IARIA/Data Analytics, Barcelona, Spain, 12–16 November 2017; pp. 65–72.
68. Chen, L.; Li, W.; Chen, H.; Geng, S. Detection of fake reviews: Analysis of sellers' manipulation behavior. *Sustainability* **2019**, *11*, 4802. [CrossRef]