Received: 1 February 2017 | Revised: 26 March 2017 | Accepted: 2 April 2017
DOI: 10.1002/cpe.4163

WILEY

SPECIAL ISSUE PAPER

CNN-based malicious user detection in social networks

Taekeun Hong[1] | **Chang Choi**[2] | **Juhyun Shin**[3]

[1] Software Convergence Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea

[2] Computer Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea

[3] Control and Measuring Robot Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea

Correspondence
Juhyun Shin, Control and Measuring Robot Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea.
Email: jhshinkr@chosun.ac.kr

Funding information
Chosun University

Summary

Following the advances in various smart devices, there are increasing numbers of users of social network services (SNS), which allows communication and information sharing in real time without limitations on distance or space. Although personal information leakage can occur through SNS, where an individual's personal details or online activities are leaked, and various financial crimes such as phishing and smishing are also possible, there are currently no countermeasures. Consequently, malicious activities are being conducted through messages toward the users who are in follow or friend relationships on SNS. Therefore, in this paper, we propose a method of assessing follow suggestions from users with less likelihood of committing malicious activities through an information-driven follow suggestion based on a categorical classification of interests using both the images and text of user posts. We ensure the objectiveness of interest categories by defining these based on DMOZ, which is established by the Open Directory Project. The images and text are learnt using a convolutional neural network, which is one of the machine learning techniques developed with a biological inspiration, and the interests are classified into categories. Users with a large number of posts are defined as certified users, and a database of certified users is established. Users with similar interests are classified, and the similarity distances between certified users and users are measured, and a follow suggestion is generated to the certified user with the most similar interest. Using the method proposed in this paper to classify the interest categories of certified users and users, precisions of 80% and 79.8% were obtained, respectively, and the overall precision was 79.93%, indicating a good classification performance overall. It is expected that the method proposed in this paper can be used to provide follow suggestions of users with less likelihood of malicious activities based on the information posted by the user.

KEYWORDS

convolutional neural network, malicious user detection, Open Directory Project, SNS Security, social network service

1 | INTRODUCTION

A social network service (SNS) is a cyberspace wherein the users can communicate, network, and share information with one another real time without limitations on distance or space. With the rapid increase in the size of the market for smart devices, the number of smart device owners has also grown, along with the number of SNS users.[1,2] With the increasing number of users, there are frequent cases of cybercrimes through SNS, including leakage of personal information such as an individual's personal details or online activities and financial crimes such as phishing and smishing.[3-6] There are cases of privacy invasion owing to personal information leakage and cases of users who suffer from unconditional criticism and malicious comments owing to unfounded allegations. Moreover, although phishing and smishing attacks had previously been performed through blogs and SMS messages, they have evolved to include SNS posts or messages, affecting a significant number of people. Therefore, in this paper, we propose a method of assessing follow suggestions of users that are less likely to conduct malicious activities based on interest categories, using both text and images of posts and a convolutional neural network (CNN), which is one of the deep learning techniques developed with a biological inspiration. This paper is organized as follows. In Section 2, we examine the CNN, which forms the theoretical basis of this study, as well as previous works on interest classification. In Section 3, we describe the algorithm for the classification of interest categories proposed in this paper. In Section 4, we describe the experiment and the results. Section 5 suggests areas for future research and concludes the paper.

2 | RELATED WORK

2.1 | Previous work on interest classification

Following the advances in and increased numbers of users of smart devices, there is an increased number of SNS users that can conveniently use SNS via smart devices. The increase in the number of SNS users, in turn, has led to an increase in the number of posts on SNS. Various studies are actively being conducted based on this trend, with interest classification being one of the topics based on SNS. Interest classification refers to the study of techniques of using a vast number of various SNS posts to analyze the individual characteristics or interests of a user and to recommend content that matches the individual interests.[7-11] In Kim et al,[7] classification criteria based on an emotion model were developed from Thayer's model, and the posts relating to movies were collected from SNS. Subsequently, the emotion perceived by the public was analyzed to measure the similarities. The Thayer's model is a 2-dimensional coordinates that represent human emotions as positive and negative. In Hong et al,[8] the latent Dirichlet allocation (LDA) algorithm was applied to the text in SNS posts of a user, and the topics were extracted based on the probability distribution of vocabulary use to classify the interests of the user. In Hong and Shin,[9] the interests of a user were classified by using both the text and images within SNS posts. Most previous works performed interest classification using text only and, hence, did not make use of all of the data from a post. Moreover, there were no criteria for classifying the interests. Therefore, in this paper, we propose a method for classifying the interests of a user using not only the text but also the images from SNS posts and describe the method of recommending users to follow that share the most similar interests by measuring the similarity distance between users.

2.2 | Convolutional neural network

CNN is one of numerous machine learning techniques and is specialized for feature extraction and perception based on neural networks. CNN was inspired by an observation in the 1950s that only part of a cat's brain is activated when an object is perceived.[12,13] Although the concept of CNN was first introduced in the 1980s as a method of computer recognition of human handwriting, there were difficulties in implementation.[14] Despite subsequent improvements through generalizations and simplifications,[15,16] it was still difficult to use CNN because of the lack of computing power. However, with improvements in computing performance and the importance of big data processing being highlighted, machine learning techniques including CNN are now receiving considerable attention.[17,18] CNN operates in a similar manner to that when an actual brain perceives an image. In the same way that a brain perceives an object, an image is divided into small regions, and the features of each region are learnt to classify the input image. CNN exhibits different structures and performance depending on the implementation, and a basic structure of a CNN is shown in Figure 1.

A CNN largely consists of a convolution layer, a pooling layer, and a fully connected (FC) layer. The feature of a part of an image is extracted in the convolution layer using a filter, and the result is subjected to the process of conversion to a channel. The size of the initial image can vary depending on the size of the input image, and the size of an image can be represented as width × height × depth. Features are calculated using Function 1.

$$\mathrm{H}(\mathrm{X}) = WX + b. \tag{1}$$

Function 1 is a formula for calculating the hypothesis in a multivariate linear regression and is used to calculate the features in a CNN. Here, X represents a set or matrix having x as elements and is a constant value that corresponds to the image; W is a set or a matrix having w as elements and corresponds to the filter. Channels are created using the filter. The pooling layer is also referred to as sampling or resizing and is used to reduce the size by extracting only the features from a channel through max pooling.[19] The FC layer, as the name suggests, is connected to all previous layers and assumes the role of finally classifying the images. Using the high-level features extracted in the previous layers, the final classification result is determined using the softmax function. In Figure 1, it is shown that the image is classified as an image of a dog, cat, boat, or bird in the FC layer.

2.3 | DMOZ

DMOZ is a project being maintained by the multinational Open Directory Project as a directory of World Wide Web links, which comprises categories in the form of directories that are established through the participation of various users consisting of voluntary editors from around the world.[20] DMOZ consists of 16 categories: Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, Kids & Teens, and DMOZ around the World. DMOZ is organized in the form of directories, with a tree structure in a hierarchy from the uppermost to the lowermost directories akin to an

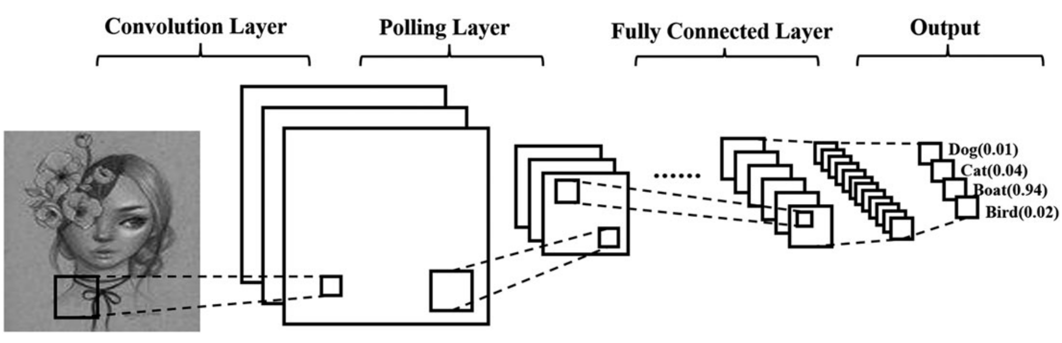

FIGURE 1 Basic structure of a convolutional neural network

actual directory. Moreover, relevant categories are also indicated. Figure 2 shows the directory structure provided by DMOZ.

It was thought that these categories formed the basis of objective categories that reflect human thoughts as these are created through the direct participation of web users. Therefore, in this paper, we redefine the categories of user interests as the categories provided by DMOZ and use these categories as the classification criteria.

3 | FOLLOW SUGGESTIONS BASED ON INTEREST CATEGORY CLASSIFICATION

In this paper, we propose a method of categorically classifying interests based on DMOZ categories through CNN using both texts and images in SNS posts and recommending users to follow with similar interests.

3.1 | System architecture

Figure 3 shows an overall system architecture of the follow suggestion method based on interest category classification proposed in this paper.

The follow suggestion method based on interest category classification proposed in this paper largely consists of training the CNN with images and text from SNS posts to be learnt, establishing a database of

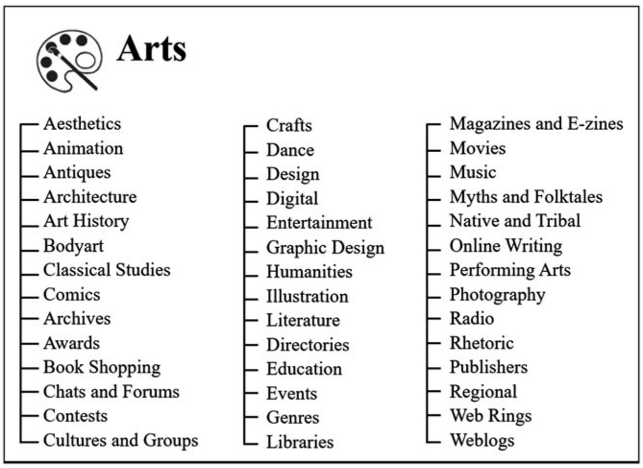

FIGURE 2 Structure of DMOZ categories

certified users, and recommending users to follow by classifying the interest category of the user.

3.2 | Establishing a certified user database

In this paper, we define a certified user as a user that assumes the role of posting various pieces of information necessary for information sharing and delivery. Therefore, a user with more than 100 posts is regarded as a certified user. To establish a database of certified users, interest categories are defined, and the CNN is trained with images and text from SNS posts. The certified user database is established by providing the images and texts in the SNS posts of an certified user as inputs to the model established through training and classifying the interest category. The certified user is assigned with an ID, and the value deduced during the classification of images and texts is stored in the certified user database.

3.2.1 | Definition of interest categories

DMOZ consists of a total of 16 categories, Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, Kids & Teens, and DMOZ around the World. Because Kids & Teens is in the form of combined words, it is replaced with Kids for simplicity. Moreover, because DMOZ around the World can be interpreted as having the same meaning as World, it is redefined as World and included in the interest categories. Table 1 shows the interest categories drawn by redefining the uppermost categories of DMOZ.

3.2.2 | CNN-based classification of interest categories

Classification of interest categories based on images

To classify the interest categories based on images, we used the stages of collecting the training images, establishing a CNN and creating a training model, collecting the images from certified users, and classifying using the CNN learning model. Instagram is used to collect the training images. Because Instagram is an SNS with an emphasis on photos, a wide variety of image information is provided. From Instagram, 1000 images are collected for each interest category

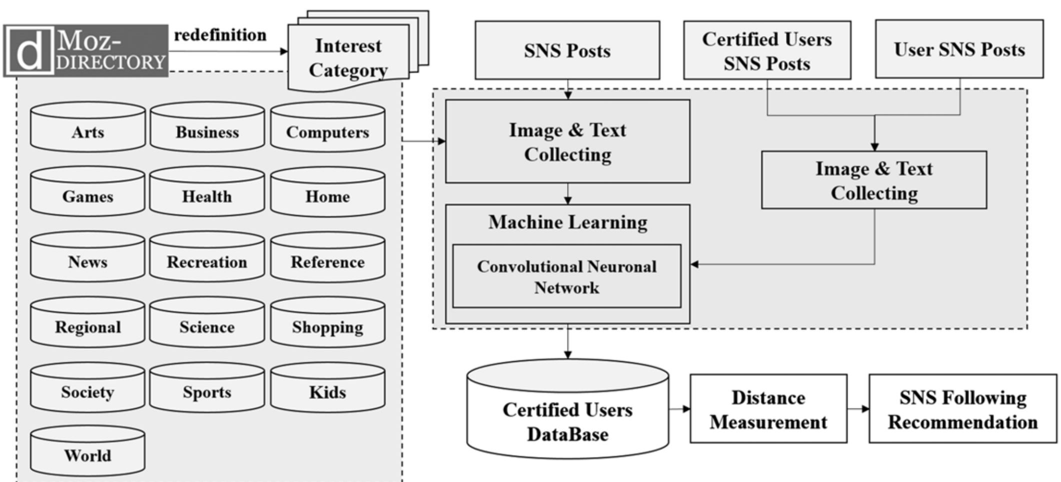

FIGURE 3 Overall system architecture, SNS, social network service

TABLE 1 Interest categories

No.	Category	No.	Category
1	Arts	9	Reference
2	Business	10	Regional
3	Computers	11	Science
4	Games	12	Shopping
5	Health	13	Society
6	Home	14	Sports
7	News	15	Kids
8	Recreation	16	World

through a hashtag search based on the interest categories. The CNN was established by modifying the Inception-v3 model provided in TensorFlow. The Inception-v3 model is a type of GoogLeNet, the CNN model used when Google won ILSVRC 2014, and resolves the problem of the amount of computation that occurs with increasing depth of layers by using filters with varying sizes.[21,22] While we follow the basic structure of Inception-v3, we create the training model by training with images collected from Instagram to learn the image data suitable for a SNS environment and allow the final classification result to be deduced as an interest category.

Figure 4 shows a flowchart of the CNN model for image classification. The conv node represents the convolution layer, and the pool node represents the pooling layer. The mixed node exhibits the basic structure of Inception-v3. Interests were classified based on the interest categories by providing the images from the users that meet the criteria to be certified users as inputs to the trained model. A total of 10 certified users were selected for each category, and 20 images were collected from each certified user, resulting in a total of 3200 images being classified. Table 2 shows an example of classifying interest categories.

Classification of interest categories based on text

To classify the interest category from text, we used the stages of collecting the training text, establishing a CNN and creating a training

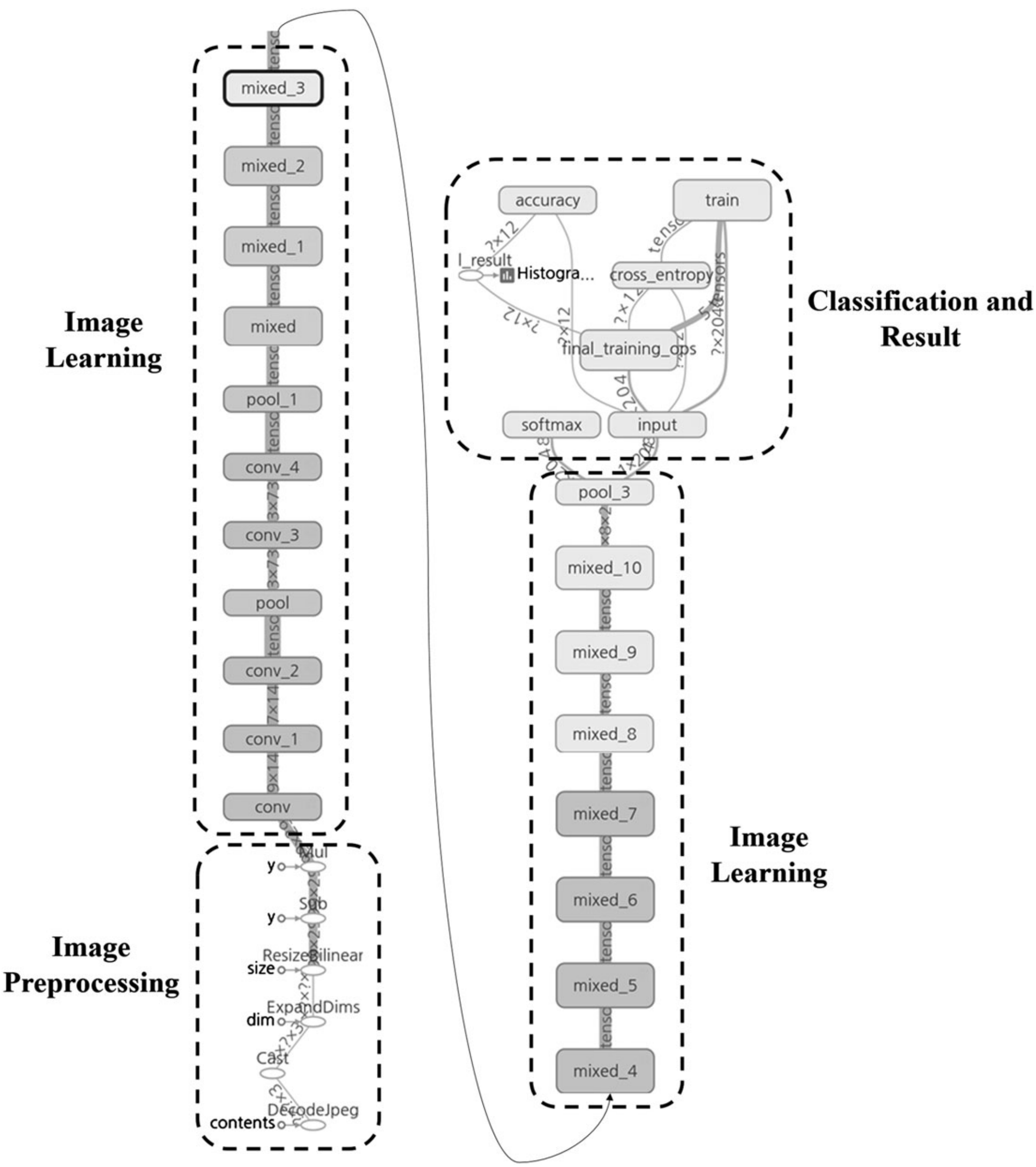

FIGURE 4 Flowchart of the convolutional neural network model for image classification

TABLE 2 Example of a resulting image classified using the trained model

Category	Score	Category	Score	Category	Score
Arts	0.22347	Society	0.23009	Society	0.23009
Shopping	0.45941	Computers	0.37283	Computers	0.37283
Arts	0.16706	Computers	0.19951	Computers	0.19951
Society	0.25665	News	0.32897	News	0.32897
Arts	0.44844	Arts	0.53489	Arts	0.35489
Arts	0.55449	News	0.24908	News	0.24908
Arts	0.70059	Home	0.19294	News	0.19294
Arts	0.34349	Society	0.27683	Home	0.27683
Science	0.27067	News	0.23541	Society	0.23541
Home	0.20683	Computers	0.15690	News	0.14690
Games	0.28088	Science	0.28198	Computers	0.28198
⋮	⋮	⋮	⋮	⋮	⋮

The entries showing "⋮" in table indicate omitted data.

model, collecting text from certified users, and classifying using the CNN training model. To collect the training text, we use Facebook, which is suitable for training data. Although Facebook is an SNS with an emphasis on text, it is also possible to mix images depending on the purpose of a post. A total of 1000 sentences of texts were collected for each category by searching the posts using the interest categories as keywords. After collecting the user posts, preprocessing is performed where the posts are divided into units of sentences to classify the text, and the words that correspond to the keywords are appended to the end of each sentence. Similar to before, TensorFlow is used to establish a CNN model for learning the text. The biggest difference to the CNN model used for image classification is that an embedding layer is required for preprocessing. The embedding layer is the stage that divides the text into units of corpora suitable for training before learning the text, which is similar to setting the filter in the convolution layer of the image CNN model. It converts sentences into a matrix consisting of word units, and the subsequent processes are identical to the image CNN model.

Figure 5 shows a flowchart of the CNN model for text classification. It largely consists of an embedding node, a conv-maxpool node, and a dropout node. As mentioned previously, the sentence-wise text provided as input to the embedding node is converted into an array of word units, and the same processes as image classification are used.

The interests were classified with respect to the interest categories by using the text from users that correspond to certified users as inputs. A total of 10 certified users were selected for each category, and text was collected from 20 posts by each certified user so that a total of 3200 sentences are classified. There were a total of 28 440 training steps. Table 3 shows an example of classifying the interests using the text from the certified users.

3.2.3 | Establishing a database using interest categories

In this section, we establish a database of certified users using the interest categories of the images and text classified in the previous section. The basic information of a certified user is tagged and stored in

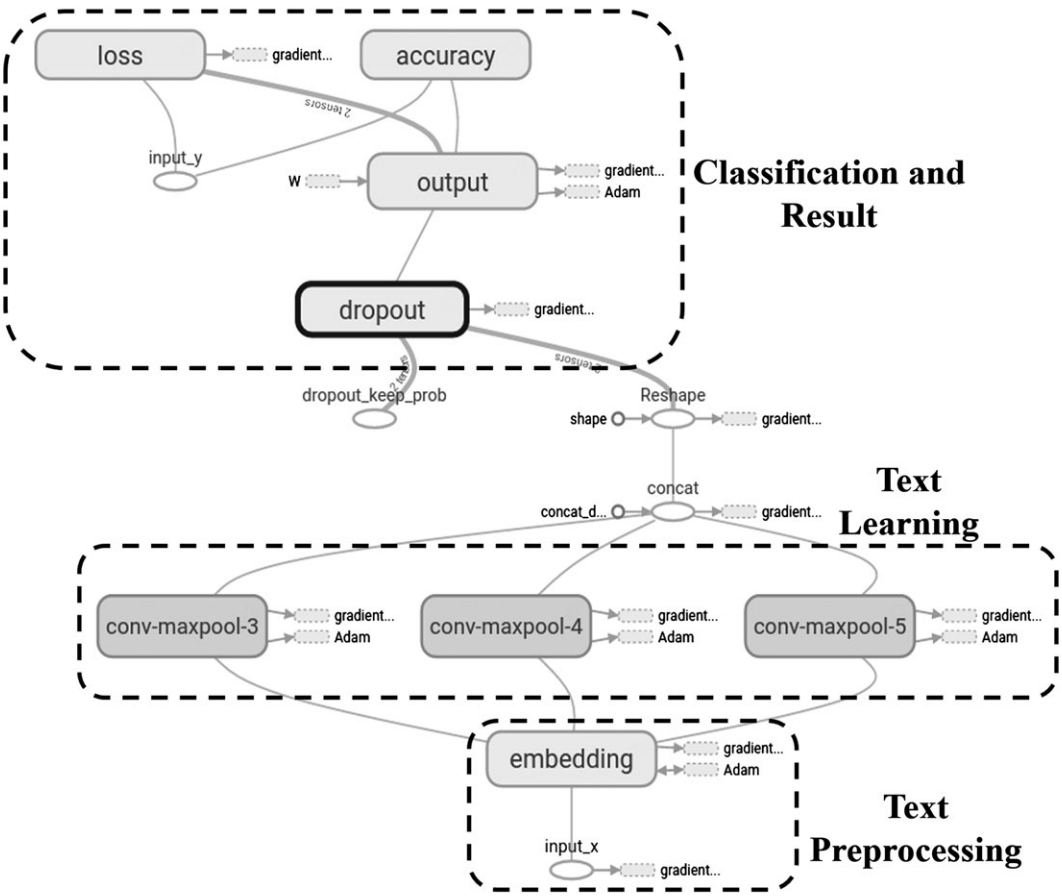

FIGURE 5 Flowchart of the convolutional neural network model for text classification

TABLE 3 Example of interest classification using the text from certified users

Category	Score	Category	Score	Category	Score
Recreation	0.00769	Recreation	0.00862	Reference	0.00724
Sports	0.00769	Arts	0.00862	Home	0.00724
Regional	0.00769	Home	0.00862	Shopping	0.01449
Business	0.00769	Shopping	0.01724	Health	0.09101
Science	0.01538	Science	0.01724	-	-
Health	0.95384	Kids	0.03448	-	-
⋮	⋮	⋮	⋮	⋮	⋮

The entries showing "-" in category and score value indicate where the classified categories of the images and text did not match.

the certified user database only when the image and text from the certified user have an identical interest category. We define the names of each interest category as ID values and assign them to the certified users, and the score values of images and text are tagged in order when the data is stored. This expresses the values of images and text from the certified user, allowing for a convenient examination of the level of values.

Table 4 illustrates the certified user database. A certified user classified as having interests in Arts and, hence, corresponding to the Arts category is assigned an ID of Arts_*n*, where *n* is an ordinal number. The parenthesized numbers in the IDs are the scores of images and text, respectively. Thus, the score values of images and text of the Arts_2 certified user in the Arts category are 0.1588 and 0.6603, respectively. Cases where the certified user ID values do not start from 1, or are not sequential and skip certain numbers, arise when the classified interest categories of an certified user do not match.

3.3 | Follow suggestions of users

In this section, we describe the method of providing follow suggestions to the user. The interest category of a user is classified, and the values of images and text are deduced in an identical manner to the classification of interest categories of certified users in Section 3. Subsequently, the certified user database is used to recommend the certified user with the most similar interests to the user among the certified users belonging to the same interest category as the user.

3.3.1 | Classification of the interest category of a user

In this section, we describe the method of classifying the interest category of a user using the images and text in the SNS posts. We classify the interest category of a user using an identical method to the classification of interests of certified users described in Section 3.2. A total of 20 images and the text extracted from 20 posts are used as inputs to classify the interests of a user. A total of 144 users are considered in classifying the interest categories, with IDs being assigned and the score values being tagged, similarly to assigning IDs and tagging scores to certified users. Table 5 shows the result of classifying the interests of users.

TABLE 4 Certified user database

Category	Contents	
Arts	Arts_2(0.1588, 0.6603) Arts_4(0.1609, 0.8915) Arts_7(0.7509, 0.8752)	Arts_3(0.7192, 0.6121) Arts_5(0.7069, 0.5983) Arts_9(0.6500, 0.9997)
Computers	Computers_1(0.3661, 0.4703) Computers_5(0.2946, 0.9085)	Computers_3(0.0868, 0.2751) Computers_8(0.1709, 0.8974)
⋮	⋮	⋮
Kids	Kids_1(0.3816, 0.8066) Kids_3(0.5665, 0.8703) Kids_6(0.5964, 0.9187) Kids_8(0.5180, 0.6500)	Kids_2(0.3404, 0.7435) Kids_4(0.4612, 0.7200) Kids_7(0.3965, 0.6741) Kids_9(0.5760, 0.6680)
World	World_1(0.4366, 0.8695) World_6(0.3096, 0.9777)	World_5(0.3952, 0.9076) World_10(0.8604, 0.9236)

TABLE 5 Classification of user interests

ID	Category	Score	ID	Category	Score
User_1	Arts	0.3856, 0.9538	User_13	-	-
User_2	Business	0.4303, 0.9747	User_15	Kids	0.4729, 0.9701
User_3	Computers	0.5597, 0.9462	⋮	⋮	⋮
User_4	Games	0.4321, 0.9548	User_136	-	-
User_5	Health	0.5564, 0.9600	User_137	Reference	0.3774, 0.9800
User_6	Home	0.4843, 0.9000	User_138	Regional	0.3830, 0.9502
User_7	News	0.6165, 0.6923	User_139	Science	0.2544, 0.9737
User_8	Recreation	0.5093, 0.9432	User_140	Shopping	0.5426, 0.9358
User_9	Reference	0.5362, 0.8621	User_141	Society	0.5103, 0.9640
User_10	Regional	0.5266, 0.9680	User_142	Sports	0.5098, 0.9640
User_11	Science	0.5445, 0.9808	User_143	Kids	0.4269, 0.9879
User_12	Shopping	0.5442, 0.9538	User_144	-	-

The entries showing "-" in category and score value indicate where the classified categories of the images and text did not match.

The user is classified only when identical image and text of interest category. Given 144 users, IDs were assigned in the form of User_*n*, with *n* being a numerical value to distinguish among the users. In the case of User_1, the category was classified as Arts, and the score values corresponding to images and texts were tagged as 0.3856 and 0.9538, respectively. The entries showing "-" in category and score value indicate where the classified categories of the images and text did not match.

Initially, the interest categories of the certified users in the certified user database and those of the user are compared to recommend the certified users belonging to the same category. The method of recommending the certified user with the most similarity is described in the following section.

3.3.2 | Follow suggestion through distance measurement

To recommend the certified user to follow showing the most similar content to the user, the distance between the user and a certified user is measured. The score values are defined as a point in a coordinate plane having *x* and *y* axes, and the similarity between the certified user and the user is measured using Euclidean distance. This can be represented using the following function.[23]

$$\text{Distance} = \sqrt{\sum_{i=1}^{n} (p^i - q^i)}. \qquad (2)$$

If 2 points *P* and *Q* have the coordinate values of $Q = (q_1, q_2, q_3 \cdots, q_n)$ and $P = (p_1, p_2, p_3 \cdots, p_n)$, the distance between the 2 points can be calculated using Function 2. Euclidean distance is often used to examine the similarity between objects having multiple properties. Multidimensional Euclidean space can be defined based on the number of properties and allows the measurement of distance. To produce the follow suggestions when the interests are the most similar even among the certified users and users in the same interest category, Function 2 is used to measure the similarity between certified users and users within the same interest category. The result of such measurement of similarity distance between certified users and users is shown in Table 6.

Table 7 shows the distance values between tag values of the images and text of certified users within the certified user database and the tag values of users calculated using Function 2. There are 6

TABLE 6 Similarity distance between certified users and users

Category	Certified user ID	User ID	Distance Value
Arts	Arts_2	User_1	0.3709
		User_17	0.4192
		User_33	0.2988
		⋮	⋮
	Arts_3	User_1	0.4775
		User_17	0.3665
		User_33	0.5500
		⋮	⋮
	Arts_4	User_1	**0.2332**
		User_17	0.3385
		User_33	**0.1220**
		⋮	⋮
	Arts_5	User_1	0.3737
		User_17	0.3706
		User_33	0.5484
		⋮	⋮
	Arts_7	User_1	0.4801
		User_17	0.2535
		User_33	0.4801
		⋮	⋮
	Arts_9	User_1	0.2683
		User_17	**0.1781**
		User_33	0.3808
		⋮	⋮
⋮	⋮	⋮	⋮
World	World_5	User_96	0.1231
		User_112	0.3124
		-	-
		⋮	⋮
	World_6	User_96	0.2012
		User_112	0.4045
		-	-
		⋮	⋮
	World_10	User_96	0.2040
		User_112	0.1539
		-	-
		⋮	⋮

The entries showing "bold text" in table indicate following recommendable case between User and Certified user. The entries showing "-" indicate where the classified Users interest did not match.

certified users in the Arts category in certified users database, who are Arts_2, Arts_3, Arts_4, Arts_5, Arts_7, and Arts_9. And 3 users in the Arts category, who are User_1, User_17, and User_33. Follow suggestions are made when the closest distance is deduced from

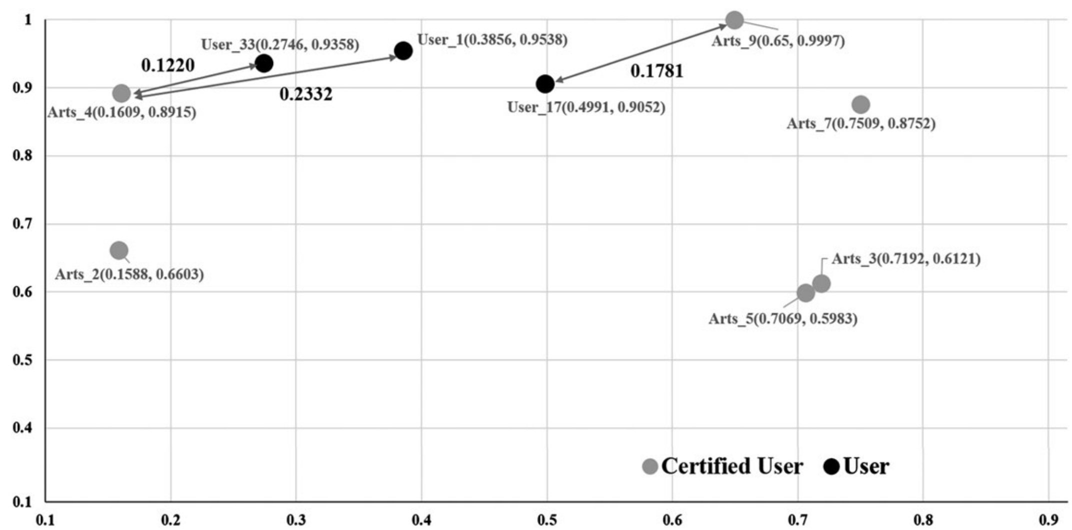

FIGURE 6 Result of similarity distance measurement between certified users and users

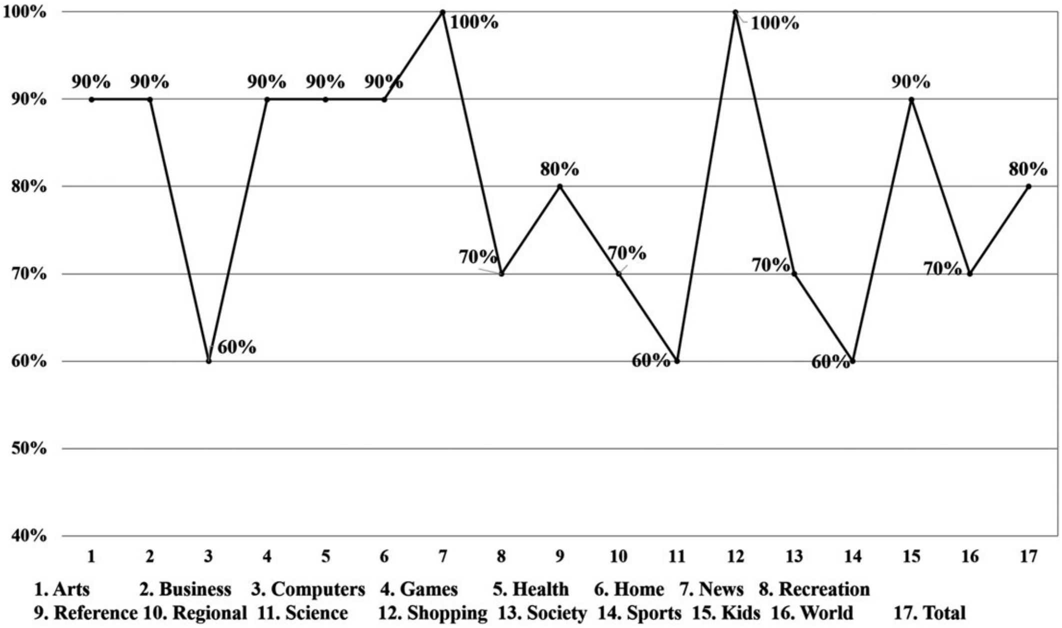

FIGURE 7 Performance of category classification for certified users

measuring similarity distances between all certified users and users. Follow suggestions can be made between Arts_4 certified user and User_1 and User_33 users, and between Arts_9 certified user and User_17 user. Figure 6 illustrates the cases where follow suggestions are possible based on the measurement of similarity distances.

Although it is possible to make the follow suggestions based on interests by classifying the interests of users, it may be the case that the users have different inclinations or specific interests even within the same interest category. Therefore, in this paper, the follow suggestions are made with the certified user having the most similar interest by measuring the distance to the user. This allows for the follow suggestion to be the most suitable and to consider the users with the least likelihood of malicious activities.

4 | EXPERIMENT AND RESULTS

The experiment in this paper uses precision to evaluate the effectiveness of the proposed interest category classification, which is measured using Function 3.

$$\text{Precision} = \frac{\textit{correctly classified post}}{\textit{classified post}} \times 100. \tag{3}$$

Precision evaluates how precise the interest category classified using images and text is and measures how many posts were correctly classified among all posts considered. Previous interest classification methods used DMOZ as the interest categories similar to the method

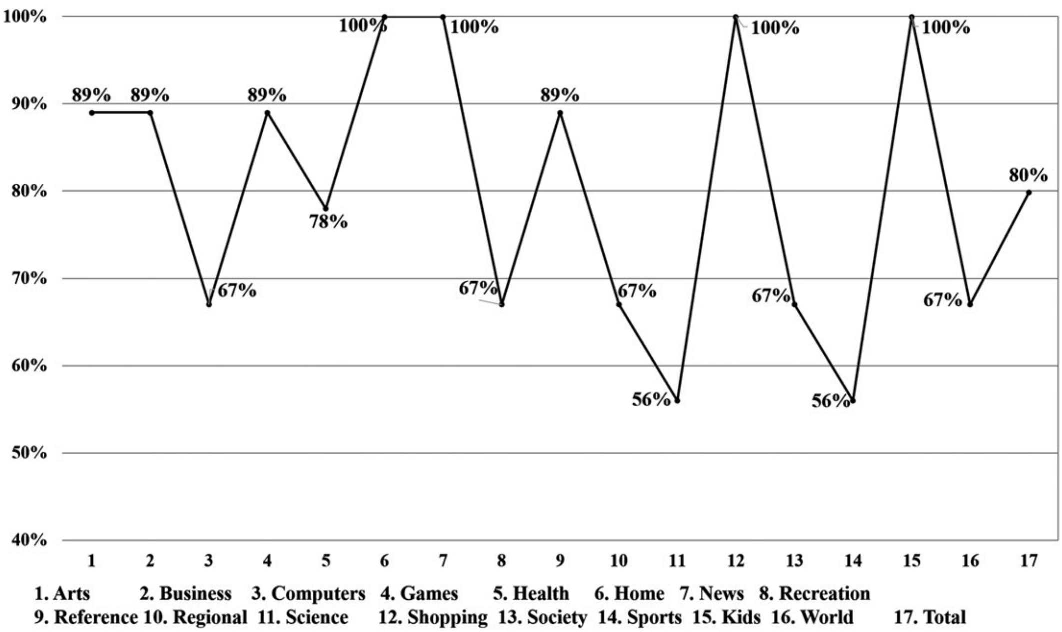

FIGURE 8 Performance of category classification for users

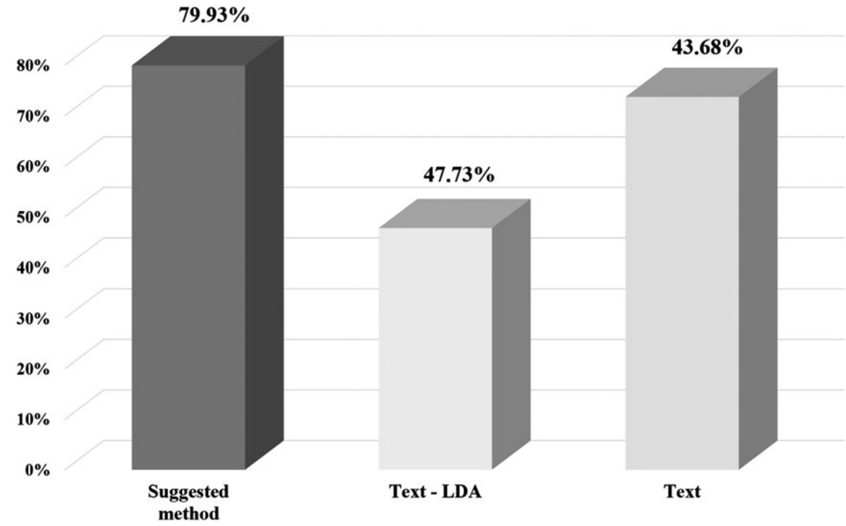

FIGURE 9 Performance comparison of the previous methods and the proposed method. LDA, latent Dirichlet allocation

proposed in this paper. We verify the performance of the method proposed in this paper by comparing against the precision measured from using the frequency of text and using LDA on text.[8,9] When the frequency of text in a post is used to classify the interest, a preprocessing step is used to eliminate the stop words, the words are extracted, and the frequency is measured. Subsequently, the 5 most frequent words are compared with DMOZ to classify the interest.[9] When LDA is used for interest classification, stop words are eliminated through a preprocessing step, the topic is classified using the LDA algorithm, which is then compared with DMOZ to classify the interest.[8] The number of posts classified in Function 3 is equal to the number of certified users and users considered for classification, and the number of correctly classified posts is equal to the number of correctly classified certified users and users. The performance of interest category classification evaluated using this method is illustrated in Figures 7, 8, and 9.

Figures 7 and 8 show the performance of the interest category classification method proposed in this paper. Figure 11 shows the precision of the interest category classification for certified users, where an overall precision of 80% is observed. Figure 8 shows the precision of the interest category classification for users, where the overall precision can be seen to be 79.86%.

Figure 9 shows the precision of both the proposed method and the previous interest classification methods. When only the frequency of text in SNS posts is used, a precision of 73.33% is observed. When the LDA algorithm is used, which extracts a topic from a document, a precision of 47.73% is observed. However, when the proposed method is used, a precision of 79.93% is obtained.

5 | CONCLUSIONS

The results of this paper allow a categorical classification of user interests based on the collection of images and text from SNS posts and learning based on CNN and provide follow suggestions to certified users that post a large amount of information. Moreover, based on the information-driven criteria, the follow suggestions can be made with users that present less of a threat of personal information leakage, phishing, or smishing, which can occur because of indiscriminate following. Previous works pertinent to SNS and follow suggestions used images and text separately, which presented difficulties in accurate identification of the intent of users. Moreover, the follow suggestions were provided based on relationships, which is not suitable for users that seek to share or acquire information. Therefore, in this paper, we used the images and text within a post simultaneously, which allowed accurate identification of the intent of users. Moreover, to define a set of clear criteria for interest classification, we used DMOZ as interest categories, which is established by the Open Directory Project as a directory where humans participate directly, achieving improved objectiveness. Furthermore, we proposed a method of recommending certified users to follow with the most similar interests by using CNN, one of the machine learning techniques, to learn and classify images and text, and measuring the distance between users and certified users belonging to the same category. From the experimental results on the proposed method, the precision of the overall interest category classification for certified users and the users was found to be 80% and 79.86%, respectively, exhibiting an overall precision of 79.93%. The categories with relatively high precision were Kids, Shopping, News, and Home. It is thought that the categories with low precision are because of similar or implied meanings of the categories, which led to the inclusion of a large amount of similar image and text data. It is expected that the proposed method can be used in user-driven services that recommend suitable information given the interests of a user, SNS marketing, and the systems that recommend users to follow with less likelihood of malicious activities or cybercrimes on SNS.

ACKNOWLEDGMENTS

This study was supported by research fund from Chosun University, 2017.

REFERENCES

1. Noh YW, Kim DY, Han JE, *et al.* Hot topic prediction scheme considering user influences in social networks. *J Korea Contents Soc*. 2015;15(8):24–36.

2. Kim WH. SNS usage trend and utilization behavior analysis. KISDI STAT report 2016; (16-07):

3. Choi JH, Sung W, Chang C, Kim PK. Personal information leakage detection method using the inference-based access control model on the Android platform. *Pervasive and Mobile Computing*. 2015;24:138–149.
4. Lee OJ, Nguyen HL, Jung JJ. Towards ontological approach on trust-aware ambient services. *IEEE Access*. 2017;5:1589–1599.
5. Hong MS, Jung JJ, Camacho D. GRSAT: a novel method on group recommendation by social affinity and trustworthiness. *Cybernetics and Systems*. 2017;48(3):140–161.
6. Jung JJ. Measuring trustworthiness of information diffusion by risk discovery process in social networking services. *Quality & Quantity*. 2014;48(3):1325–1336.
7. Kim KM, Kim DY, Lee JH. Measuring similarity between movies based on polarity of tweets. *J Korean Instit Intel Syst*. 2014;24(3):292–297.
8. Hong TE, Ju YJ, Shin JH. The method of SNS following recommendation using topic and category. *Proceedings of Conference on KOREA Multimedia Society*. 2016;19(1):89–92.
9. Hong TE, Shin JH. Recommendation method of SNS following to category classification of image and text information. *J Smart Media*. 2016;5(3):54–61.
10. Ogiela L, Ogiela MR. Data mining and semantic inference in cognitive systems. 2014 International Conference on Intelligent Networking and Collaborative System(IEEE INCoS 2014), SEP 2014. Italy, 2014; 257-261.
11. Choi JH, Choi C, Lee EJ, Kim PK. Markov logic network based social relation inference for personalized social search. In: *New Trends in Computational Collective Intelligence*. Vol. 572; 2015:195–202.
12. Hubel D, Wiesel T. Receptive fields of single neurones in the cat's striate cortex. *J Physiol*. 1959;143(3):574–591.
13. Hubel D, Wiesel T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex 1962; 160(1): 106-154.
14. LeCun Y, Boser B, Denker J, Henderson D. Howard. R, Hubbard W, Jackel D. Backpropagation applied to handwritten zip code recognition. *J Neural Comput*. 1989;1(4):541–551.
15. Behnke S. *Hierarchical Neural Networks for Image Interpretation*. Berlin Heidelberg: Springer-Verlag; 2003.
16. Simard P, Steinkraus D, Platt J. Best practices for convolutional neural networks applied to visual document analysis. ICDAR 2003; 2: 958.
17. Ahn SM. Deep learning architectures and applications. *J Intel Info Syst*. 2016;22(2):127–142.
18. Ogiela L, Ogiela MR. Advances in cognitive information systems. *Cognit Syst Monogr*. 2012;17:1–18.
19. Jang HW, Cho SS. Automatic tagging for social images using convolution neural networks. *J Korea Info Sci Soc*. 2016;43(1):47–53.
20. DMOZ. http://www.dmoz.org/ [20 november 2016].
21. Szegedy C, Vanhoucke V, Loffe S, *et al*. Rethinking the inception architecture for computer vision. *J Comp Vision and Pattern Recog*. 2015;
22. Tensorflow-inception, serving inception model with TensorFlow serving and Kubernetes.
23. Yun JC, Kim HC, Kim JJ, Youn SD. A study on the CBR pattern using similarity and the Euclidean calculation pattern. *J Korea Instit Info Comm Eng*. 2010;14(4):875–885.

How to cite this article: Hong T, Choi C, Shin J. CNN-based malicious user detection in social networks. *Concurrency Computat: Pract Exper*. 2018;30:e4163. https://doi.org/10.1002/cpe.4163