Non-Newsworthy Message Removal for Efficient Credibility Assessment

Chaluemwut Noyunsan, Tatpong Katanyukul, and Kanda Runapongsa Saikaew

***Abstract*—Social media has received overwhelming interest globally from their functions to post users contents, some of which has not been validated and may contain false or misleading content. There have been several studies to assess the credibility of social media posts to tackle such problem. Most existing online assessment systems evaluated the credibility of every post. This practice may be suboptimal. Many contents are not newsworthy (e.g., selfies and personal opinions, which are irrelevant to credibility notion). Assigning credibility score to a non-newsworthy post may confuse users. In addition, a recent study has shown that the inclusion of such irrelevant non-newsworthy data deteriorates the quality of credibility assessment. Therefore, identification and exclusion of non-newsworthy posts are crucial to reliable credibility assessment. This article investigates how different types of post features are effective for automatic non-newsworthiness removal. Three post features, i.e., text, topic, and social features, were evaluated with two classification methods which were machine learning and cosine similarity. Our findings reveal the essence of social features and its combination for non-newsworthiness identification.**

***Index Terms*—Social network analysis, credibility measurement, information credibility, supervised machine learning, TF-IDF.**

I. Introduction

Social media has been popular worldwide. It provides user-generated contents. There are several types of contents, such as messages, pictures, videos, and live streams. Facebook serves various contents, e.g., messages, pictures, and URLs. Twitter is a microblogging service on which users can post short messages (limited to 140 characters). Youtube is a video-oriented service. By April 2016, Facebook, QQ (China social media), Instagram, and Twitter had 1.5 billion, 853 million, 400 million, and 320 million monthly active users, respectively[1]. Thus, social media analysis attracts wide interests and active research studies [1], [2].

Castillo *et al.* and Gupta *et al.* investigated approaches to verify trust ability of social media posts [3], [4]. They defined credibility scores as a means for verification. Credibility score indicates a trustworthiness level. A user is supposed to be able to trust posts with high credibility ratings. Several studies [4], [5] assigned credibility scores to all social media posts. However, credibility score on a non-newsworthy post may confuse a user or cause loss of faith in the credibility assessment system. In terms of computing, it would take a longer time to process the data. Moreover, Noyunsan *et al.* [6] showed that inclusion of non-newsworthy posts in building a learning-based credibility assessment lead to poorer quality of a prediction model. Therefore, identification and exclusion of non-newsworthy posts were essential to reliable and effective credibility assessment.

Non-newsworthy posts are defined as unimportant posts. That is, the posts are not useful or interesting to the general public. Such posts can be categorized as non-informative posts, such as selfies and personal matters. These kinds of posts affect only a small circle of persons. Putting credibility score on non informative posts may confuse a user. Excluding this kind of posts from credibility assessment will enhance the total user satisfaction and improve an overall credibility assessment [6]. It will also lighten the computing load as unnecessary posts are discarded. Note that [3] and [4] reported that non-newsworthy posts comprised about 40% and 56% of all online posts in 2011 and 2013, respectively.

Social network analysis has been an attractive and active research topic in recent years due to the widespread daily usage of social media networking. Imran *et al.* [7] surveyed several techniques, such as filtering, classifying, ranking, and summarizing for processing social media data in an emergency case. Mitra *et al.* [8] created a large credibility corpus (called CREDBANK) which had 60 million tweets out of 1,049 real-world events by using human annotators.

Recently, automatic credibility assessment has been explored in social media [3], [7], [9]-[11]. Most studies worked on Twitter because data could be quickly retrieved via Twitter APIs. These studies retrieved data from Twitter APIs and employed humans to label the data [3], [4]. After that, data were separated into training data sets and test data sets. Next, a model was created from training data, and the performance of the model was measured by using test data sets.

In one of the earliest works, Castillo *et al.* [3] developed an automatic credibility measurement in Twitter. They grouped Twitter features into four groups which are Message, User, Topic, and Propagation. J48 decision tree was used for classification. Their proposed method achieved 70%-80% precision and recall according to user feedbacks. Gupta *et al.* [9] investigated on computing the credibility of high impact events, such as "Hurricane Irene", "Google acquires Motorola Mobility", and "New Facebook Messenger". They used Pseudo Relevance Feedback for improving prediction efficiency.

Aggarwal *et al.* [12] attempted to find tweets with phishing content. They used training data from URL online blacklist services, such as PhishTank[2] and Google Safebrowsin[3].

Manuscript received August 25, 2017; revised November 29, 2017.

The authors are with Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002 Thailand (e-mail: chaluemwut@hotmail.com, tatpong@kku.ac.th, krunapon@kku.ac.th).

[1] http://www.statista.com/statistics/272014/global-social-networks-ranked- by-number-of-users/

[2] http://www.phishtank.com/

When tweets were found in the blacklist service, they were labeled as phishing as opposed to being labeled as safe. Aggarwal *et al.* compared three algorithms which were naive Bayes, decision trees, and random forest. They found that random forest was the best performing method.

Gupta *et al.* [4] created the prediction model and assessed online contents in real-time and developed a Google Chrome extension, called Tweetcred, which could present credibility score on Twitter timeline in real-time. Tweets were acquired with Twitter APIs to prepare training data. Humans were employed to label acquired tweets for credibility scores, on a scale of 1-7 (1 for the lowest credibility and 7 for the highest credibility). SVM-rank was used as a prediction model.

Assessment of only newsworthy posts has been studied in the context of social media data summarization. Canneyt *et al.* [13] summarized messages in Twitter. They detected whether tweets were newsworthy before executing summa- rization process. Removal of non-newsworthy posts before prediction process could improve credibility assessment performance [6]. They added non-newsworthy posts in training data and found that the larger non-newsworthy posts the poorer the credibility assessment performance. The relationship between the number of non-newsworthy posts and the performance degradation of credibility assessment was in a quadratic equation. In the context of social media data credibility, several previous works [3], [4] studied non-newsworthy posts, but none had attempted to remove non-newsworthy posts systematically and automatically. They removed non-newsworthy posts by using a manual process. Castillo *et al.* [3] manually filtered out non-newsworthy posts before performing the experiment. Gupta *et al.* [4] used only newsworthy posts, removed non-newsworthiness by manual process, in the training data to build a model but they included all posts in the test data. Castillo *et al.* [3] and Gupta *et al.* [4] removed non-newsworthy posts by using worker label data to news-worthiness or non-newsworthiness.

Unlike other research works, this article presents the design, implementation, and evaluation of the system that automatically removes non-newsworthy posts. Our work aims to investigate an approach to build an effective non-newsworthiness identification system. We built automatic non-newsworthy post removal systems which were developed based on text, topic, and post features. Three different types of these systems were processed using two classification algorithms, i.e., cosine similarity-based and random forest methods. Cosine similarity was a method used for topic detection using text classification. Random forest was chosen as a supervised machine learning algorithm since it has been used widely for classification. Social media features were features which users had interactive activities with posts. Text features were features which a message in the post was processed.

There were several research studies which detected interesting posts [2], [14] . Yang & Rim (2014) defined interesting posts as posts that were possibly being "of potential interest to not only the authors and their followers but a wider audience" and uninteresting as being "only interesting to the authors and their friends due to personal interests." However, they associated "general and mundane topics appear any time spans" to uninteresting posts. This notion is clearly seen in the development of their system, but it sets their perception of interestingness apart from our newsworthiness. For example, common mundane news, e.g., missing child, may appear to be uninteresting, but it is newsworthy. Non-newsworthy posts are the posts that are unimportant or have little impact on other people. Non-newsworthy posts do not need credibility measurement because it may make user confuse and make the prediction performance drop. This article studies the impact of non-newsworthiness with credibility measurement. A non-newsworthy post is the post conveying information that is of no effect and of no interest to general public.

The rest of the article is organized as follows: a literature review is summarized and analyzed in Section II, methodology is explained in Section III, experimental results are described in Section IV. Finally, the conclusions of the study are drawn and future work is discussed in Section V.

II. Methodology

Fig. 1 shows the system overview. To prepare the system, firstly, we collected credibility training data by crowdsourcing. Via crowdsourcing, human annotators labeled whether a Facebook post was newsworthy or not.

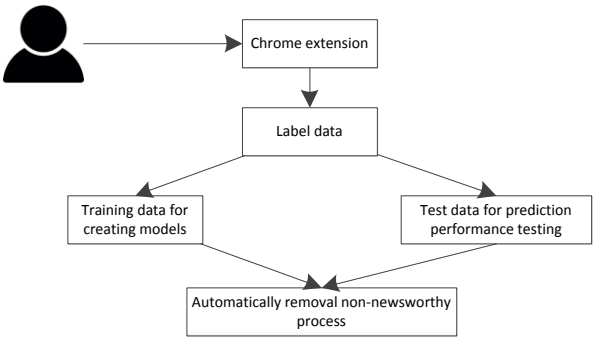

Fig. 1. The system overview.

Via crowdsourcing, human annotators labeled whether a Facebook post was newsworthy or not. The features of each post can be divided into three parts including: social media features, text features, and topic features.

A. Data collection

Firstly, we collected Facebook data by implementing a Google Chrome extension, which was called FBNewsworthEvaluation. Through FBNewsworthyEvaluat ion installation, we collected Facebook features, such as the number of likes, and the number of comments. Table II shows all Facebook features employed in our system. There were 4,436 records collected between January 20, 2017 and February 8, 2017. Fig. 2. shows the screenshot of a Facebook post on Google Chrome after FBNewsworthEvaluation installation. For each post, a user was asked to judge its newsworthiness. Users' feedbacks along with the post details were sent to our server for data collection.

B. Removal of Non-Newsworthy Posts

This article uses cosine similarity and machine learning to remove non-newsworthy posts. Cosine similarity is widely used in Topic Detection and Tracking (TDT) research [15] when focuses on searching event-based organization in

[3] https://developers.google.com/safe- browsing/

broadcast media. Machine learning can solve problems in several perspectives, such as control robot, computer vision, and data mining. There are various types of machine learning algorithms, such as supervised machine learning algorithms, unsupervised machine learning algorithms, and reinforcement learning algorithms.

Fig. 2. A Facebook post on Google Chrome with FBNewsworthEvaluation Extension.

1) Cosine similarity is the measurement of the similarity between two TF-IDF[4] vectors. It is a value of angles between two vectors (possible value is [0,1]). Cosine similarity between $\overrightarrow{d_1}$and $\overrightarrow{d_2}$ can be computed using Equation [16]

$$cos(\theta) = \frac{\overrightarrow{d_1}\cdot\overrightarrow{d_2}}{|\overrightarrow{d_1}|\cdot|\overrightarrow{d_2}|}. \quad (1)$$

TF-IDF is a method vectoring text-based documents as a vector of real numbers. Each real number, denoted by β, represents a discounted frequency of a particular word. The discount factor reflects how uncommon a particular word is in the corpus. Given corpus M consisting N messages, m_i, m_2 ,.., m_N ,messages. TF-IDF value of word w_j in m_i message can be computed by Equation 2:

$$\beta_j = \frac{f(m_i\, w_j)}{\sum_{t\in Q_i} f(m_i, w_k)} \cdot \log\frac{N}{\sum_{n=1}^{N}\delta(m_n, w_j)}, \quad (2)$$

where $f(m_i, w_j)$ represents a number of occurrences of word w_j in message m_i, Q_i represents a number of distinct words in message m_i, N is a number of messages in the corpus M, and $\delta(m_n, w_j) = 1$ if message m_n contains word w_j and $(m_n, w_j) = 0$ otherwise.

2) Supervised Learning Method: Supervised machine learning, or supervised learning, refers to a prediction model giving output y for input x (a set of attributes) where the quality of the prediction can be improved by relevant examples, called training data. Training data is comprised of various inputs and correct outputs. When the output, often called label, can take only one of two possible values, the task is categorized as binary classification. However, if the output can take one of multiple possible values, then the problem is categorized as multiclass classification.

We chose to implement Random forest, which is one of the most widely used supervised learning methods. Regarding implementation details, we used Scikit-learn[5], which is a commonly used Python machine learning library. Attributes or features which were used as inputs x of our decision tree model was extracted from a social media post. The features are listed in Table I.

TABLE I: POST FEATURES USED IN OUR DECISION TREE MODEL

Social media features: number of likes, number of shares, number of comments, number of hashtags, number of images, whether there is an URL, whether a post is a video post, number of friends who are tagged, whether a post is public, whether a post is shared with location, whether a post is shared with feeling status
Text features: number of words, number of question marks, number of characters, number of exclamation marks, number of words in a dictionary

III. EXPERIMENTAL RESULTS

This article focuses on the comparison of three features to identify non-newsworthy posts. Three features were used to create a classification including topic features, text features, and social features. We separated 80%, 3549 distinct records, of all data into training data while the remainder, 887 distinct records, was test data. The experiments were repeated 10 times.

In machine learning methods, the model was created from training data and test performance of the model by using test data. In cosine similarity method, TF-IDF vectors were created from training data and test data. TF-IDF of test data was compared with data in training data by using cosine similarity. We compared TF-IDF of each test data with all TF-IDF of training data. Label of training data which made the highest value of cosine similarity was determined to output label.

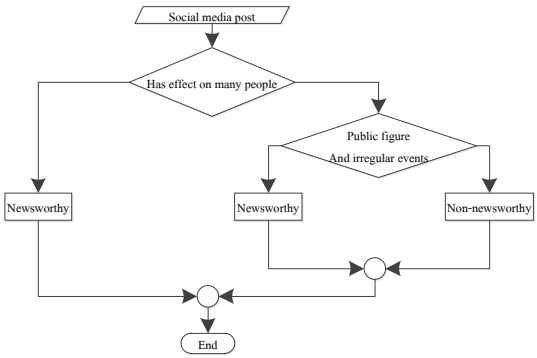

Fig. 3. How to determine non-newsworthy posts.

Fig. 3 shows the procedure for non-newsworthiness determination. Newsworthy posts are the posts that affect several people. The posts which do not affect other people are non-newsworthy posts.

F1-Score is a performance metric of binary classification. It can be computed using $F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, where $precision = \frac{TP}{TP+FP}$, and $recall = \frac{TP}{TP+FN}$, In binary classification, TP represents cases when a classifier correctly predicts a positive label (non-newsworthy label). FP represents cases when a classifier incorrectly predicts a positive label. FN represents cases when a classifier incorrectly predicts a negative label.

Fig. 4 shows the prediction performance of topic features, text features, and social features. The x-axis represents a boxplot of topic features, text features, and social features while y-axis represents F1-Score.

[4] https://en.wikipedia.org/wiki/Tf–idf
[5] http://scikit-learn.org

As shown in Fig. 4, classifiers using only social features outperform classifiers using text or topic features alone. Significance tests, shown in Table II, also confirm the superior performance using social features over using text or topic features alone.

Table III shows t-test results comparing performance using social features (SF), performance using text features (TX), performance using the combination of social and topic features (SF+TO), performance using the combination of social and text features (SF+TX), and performance using the combination of all features types (SF+TO+TX). The normality assumption is valid for all data (validated by Shapiro-Wilk test). An entry in the table represents a decision result (*W*, or *L*) and p-value (in parentheses) obtained for two-side t-test ($H_0 : \mu_1 = \mu_2$, $H_0: \mu_1 \neq \mu_2$). Decision result of 'W' means that the t-test confirms the difference (favor of H0). That is, at confidence level of 0.05 performance using social features (SF) is significantly better than the performance using topic features only (TO) and using topic and text features (TX). Decision result 'L' means that t-test confirms the difference at 0.05 and performance using social features (SF) is significantly worse than the performance using (SF+TO), (SF+TX), and (SF+TO+TX).

Social features performance which had F1-Score at 0.6 was the best. Topic and text features had quite similar performance with the results of F1-Score at 0.5. Social features, such as number of likes, number of shares, e.g., achieved higher performance than text features and topic features. Text features and topic features use features of word, such as number of words, word length, or a word vector. However, using only word characteristics in a sentence alone cannot identify newsworthiness. Newsworthiness depends on the context of the sentence such as, 'police' in sentence 'The police who escapes from the prison is going to this street' is newsworthy, and 'police' in sentence 'This police is a old man' is non-newsworthy. Social features yielded the best performance because non-newsworthy posts usually did not receive an attention from users while newsworthy posts usually were in the interest of others and thus had numbers of social features with higher values.

Fig. 4 shows the boxplots of prediction performance (F1-Score) when using only social features and when using the combinations of features (social and topic features, social and text features, and all feature types). From Fig. 6, using the combination of feature types results in higher performance than using solely social features. This is also confirmed by significance test results as shown in Table II. The combination between social features with text or topic can improve prediction performance.

Fig. 4 shows non-newsworthy performance by using cosine similarity and machine learning. When using either cosine similarity or machine learning, topic features and text features have almost the same performance. Social features and the combinations of social features and other features (text features, topic features, topic and text features) show the highest performance. Nevertheless, compared with when using cosine similarity, the combinations of social features and other features (text features, topic features, topic and text features) when using machine learning perform better. In our experiment, random forest is used as a supervised machine learning algorithm because it is precise and fast [18]. Random forest uses several random decision trees to vote the prediction output. One of the reasons why random forest can classify newsworthy posts and non-newsworthy posts better than using cosine similarity of TF-IDF probably because it can solve an over fitting problem more efficiently. Cosine similarity compares two vectors by using sizes of angle degrees but it does not consider the length of two vectors. Regarding prediction time, cosine similarity compares test data one by one with training data thus it requires a large computation time. On the other hand, random forest compares test data with the model. Such process requires less time compared with when using cosine similarity.

TABLE II: INFORMATION OF T-TEST VALUES

	Topic features (TO)	Text features (TX)	Social combination with topic features (SF+TO)	Social combination with text features (SF+TX)	Social, text, and topic combination (SF+TO+TX)
Social fea- tures (SF)	W(4e-85)	W(2e-84)	L(1e-15)	L(7e-25)	L(8e-26)

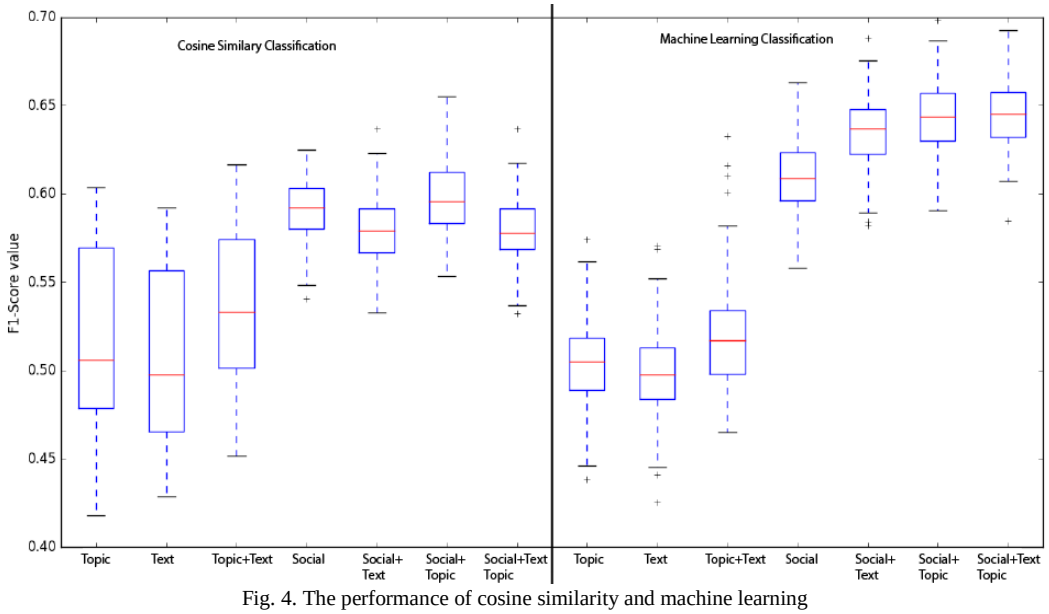

Fig. 4. The performance of cosine similarity and machine learning

In practice, topic features extraction is difficult to implement efficiently. Topic features extraction usually relies on word-vector or bag-of-words scheme that, which maps a given text to a vector of word counts for a set of predefined words. When a new message containing a word not in a predefined set, it is either the word will be ignored and not be presented in topic features or that topic features model has to be rebuilt. In addition, a number of all possible words are excessively, if not infinitely, large. Careful word selection is crucial to an effective topic features scheme. These issues render current topic features scheme difficult for practical newsworthy message identification.

IV. Conclusions

In this article, non-newsworthy removal was studied by using machine learning and cosine similarity. Text, topic, and social features were used for classification. The combination between social and other features (text and topic) resulted in the highest performance when using machine learning. Our finding reveals the viability of newsworthiness identification. This enables a more efficient credibility assessment system and would possibly make an online society a little safer.

For potential future direction, a study of non-newsworthiness removal on a real-time system would allow more thorough evaluation for factors or issues that may not be fully realized in an offline setting. Another direction that is worth to further studies is to explore an application of reinforcement learning for non-newsworthiness removal.

Regarding social features, previous studies on either trust-worthiness or newsworthiness of social media treat social features as if they were static values. However, social features, e.g., number of likes, number of comments, number of shares, vary in time. This line of research may lead to more efficient implementation of social media credibility assessment system and higher understanding of social media behaviors and its association public at large.

References

[1] C. C. Aggarwal and C. Zhai, *Mining Text Data*, Springer Science & Business Media, 2012.

[2] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar, "Detecting uninteresting content in text streams," in *Proc. SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[3] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc.* the 20th international conference on World wide web, 2011, pp. 675-684.

[4] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," in *Proc.* International Conference on Social Informatics, 2014, pp. 228-243.

[5] C. Noyunsan, T. Katanyukul, and K. R. Saikaew, "Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes," *Eng. Appl. Sci. Res*., vol. 45.

[6] C. Noyunsan, T. Katanyukul, C. K. Leung, and K. R. Saikaew, "Effects of the inclusion of non-newsworthy messages in credibility assessment," in *Proc. Mexican International Conference on Artificial Intelligence*, 2016, pp. 185-195.

[7] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, and A. Alamri, "A credibility assessment model for online social network content," *Social Data Mining and Analysis to Prediction and Community Detection*, Springer, pp. 61-77, 2017.

[8] T. Mitra and E. Gilbert, "CREDBANK: A large-scale social media corpus with associated credibility annotations," in *Proc. ICWSM*, 2015, pp. 258-267.

[9] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proc. the 1st Workshop on Privacy and Security in Online Social Media*, 2012, p. 2.

[10] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on twitter," in *Proc. 2012 SIAM International Conference on Data Mining*, 2012, pp. 153-164.

[11] T. Kawabe *et al.*, "Tweet credibility analysis evaluation by improving sentiment dictionary," in *Proc. IEEE Congress on Evolutionary Computation (CEC)*, 2015, pp. 2354-2361.

[12] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "Phishari: automatic realtime phishing detection on twitter," *eCrime Researchers Summit (eCrime)*, 2012, pp. 1-12.

[13] S. V. Canneyt, M. Feys, S. Schockaert, T. Demeester, C. Develder, and B. Dhoedt, "Detecting newsworthy topics in twitter," *Data Challenge (SNOW 2014)*, pp. 1-8, 2014.

[14] M.-C. Yang and H.-C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4330-4336, 2014.

[15] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, Topic detection and tracking pilot study final report, 1998.

[16] G. Chowdhury, Introduction to modern information retrieval, Facet publishing, 2010.

Chaluemwut Noyunsan was born in Khon Kaen, Thailand. He received the B.S. degree in computer engineering from Khon Kaen University, Thailand, in 2003, and the M.S. degrees in computer science from Chulalongkorn University, Thailand, in 2009, respectively. Currently, he is studying PhD degree in the Department of Computer Engineering, Khon Kaen University. He current research interests include social network analysis and information credibility measurement.

Tatpong Katanyukul got the B.Eng. in electronics engineering from King Mongkut Institute of Technology, Ladkrabang; M.Eng. in computer science from Asian Institute of Technology, and Ph.D. in mechanical engineering from Colorado State University. His academic areas of interest are in approximate dynamic programming, including reinforcement learning. machine learning applications.

Kanda Runapongsa Saikaew was born in Chiang Rai, Thailand. She received the B.S. degree in electrical and computer engineering from Carnegie Mellon University, Pennsylvania, USA, in 1997, the M.S. and Ph.D. degrees in computer science and engineering from the University of Michigan at Ann Arbor, in 1999 and 2003, respectively. In 2003, she joined the Department of Computer Engineering, Khon Kaen University, as a lecturer, and became an associate professor in 2015. Her current research interests include social network analysis and machine learning.