Status Analysis in Social Media: Input for Potential Business Trends

Paula Jean M. Castro
Department of Information Technology
Technological Institute of the Philippines
Quezon City, Philippines

Rosmina Joy M. Cabauatan
Department of Information Technology
Technological Institute of the Philippines
Quezon City, Philippines

Bartolome T. Tanguilig III
Office of the AVPAA
Technological Institute of the Philippines
Quezon City, Philippines

***Abstract*—Social Media such as Twitter, Facebook, and LinkedIn provides people an avenue to virtually communicate and share implicit and explicit expressions about anything that interests them. These are normally in the form of posts and likes. Regardless of who instigated the posts, it inherently invites feedbacks in the form of likes as expressions of agreement or appreciation. In this study, a corpus of these types of explicit expressions was collected using Facebook Graph Application Platform Interface. With an end-in-view of utilizing it for the identification of potential business trends that could most likely be embraced by a specific clusters of users, the data were pre-processed and clustered using K-Means data mining technique. Results show that interests of clusters of user unveil potential business trends, probable customers and insights for outstretched networked marketing strategies.**

***Keywords*—social media; data preprocessing; clustering; RapidMiner; tf-idf;**

I. INTRODUCTION

With its inherent means of creating wide array of virtual communities, social media has potentials in producing innovative business paradigms. While several perspectives could possibly arise from user information, interactions create more meaningful indicators of how these communities could incrementally evolve and what probable information could possibly be drawn as input in identifying business trends.

Most information-abounding networking sites include Facebook, Twitter and LinkedIn. With 955 million active users [2], Facebook has gained the largest. Interactions among these users are typically stimulated by their lists of favorites called Likes.

Facebook Likes, is a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases [12].

An increasing number of companies have adopted Facebook as a marketing tool [6]. As testified in the study of Surma, et al [1] popularity of social media has gained great opportunities to reach large audience through viral marketing campaigns [11]. H. Wu, et al [14] interpreted Facebook pages to generate key factors that could attract customers and for young entrepreneurs to react to new postings. Facebook is likewise used by S. Bhatia, et al [16] to monitor feedback from of customers necessary for identifying possible corrective measures.

While Facebook marketers take advantage of their network of friends to introduce products they believed to be vendible, only a few took advantage of their interactions through status and likes as prime source of understanding how to reach out potential users who could possibly matter most to the business to optimize marketing.

It is in this premise that this study could help users to identify what possible business and specific users who are most likely to become the customers.

In understanding users' network of friends and circles of interactions, clustering data mining technique has been utilized. This technique is broadly used for data analysis. From the business viewpoint, this could help marketers discover distinct groups in order to gain insights on specific products to be introduced and in a timely offer.

II. RELATED WORKS

Several business-oriented studies have been conducted using Facebook data.

The worked of K.C. Gull, et al [5] analyzed Facebook users' interests using Fuzzy c-Means Clustering algorithm to effectively improve marketing plans. A method for extracting users' interests is likewise proposed by J. Kim, et al [17] using the Facebook's social plug-in "Like". This was done by calculating term frequency of extracted nouns and Likes.

Dialogue between young entrepreneurs and their audience of Facebook pages were interpreted in the research of H. Wu, et al [14] to better understand how to post interesting topics in order to strengthen marketing communications and increase market shares. E. Pöyry, et al [6] developed a model for consumer behavior on company-hosted Facebook community page by exploring the outcomes of the users' usage behavior in terms of purchase intentions. N. Salamanos, et al [2] correlated social network communities as defined by a community detection algorithm and Facebook pages annotated as Likes by its users to examine the relation between the underlined social dynamic, as expressed indirectly by a community structure, with the users' characteristics represented by Likes.

C. Canali, et al [11] proposed a quantitative methodology that can support social network analysis for identifying relevant users using Principal Component Analysis (PCA) as qualifying point to select and combine user attributes into characteristics that are meaningful for analysis. This was found to have a marketing advantage to Internet based business as it allows one to find users who are most responsive to different types of dissemination strategies, such as viral marketing or brand-based campaign.

While most of the studies used Facebook pages and likes, this study used both status and likes to understand better the interests of users. Topics posted in Facebook pages invite implicit comments or feedbacks from followers while status and likes are explicit expressions of interests of someone who posted it. This could contribute to inherently identify the personal interest of users.

III. Methodology and Discussion

The main objective of this paper is to cluster Facebook users based on their status posts and likes in order to identify potential business opportunities for specific groups.

1. Data Collection

A total of 50 volunteers with a minimum of 210 status and 88 likes per user from year 2012 to 2014 were collected using an access token of Facebook Graph API Explorer. After removing some insignificant attributes such as photos and emoticons, a total of 14, 586 status and 4,406 likes were produced. From JavaScript Object Notation (JSON) format, the data was converted to Comma-Separated Value (CSV) and Excel formats. Figure 1 shows the processes that were carried out to determine the most liked and posted topics.

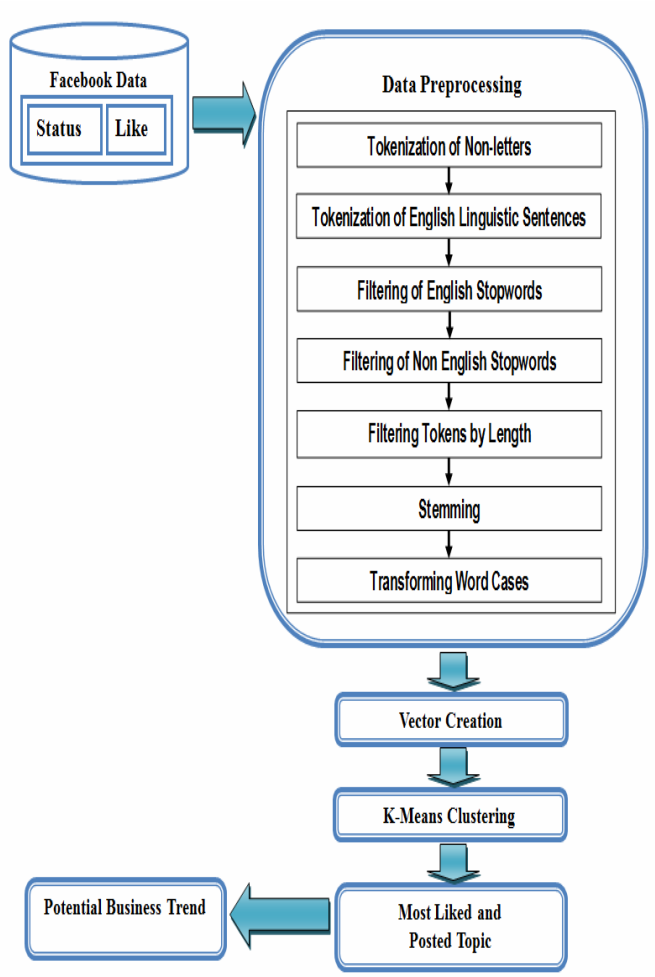

Figure 1. The Data Analysis Process

2. Data Preprocessing

To prepare the collected data for further processing, several transformations were performed using RapidMiner text processing plug-ins such as text and WordNet extensions.

Tokenization of non-letters and linguistic sentences were performed to exclude special characters or symbols. In the Filtering of English Stopwords stage, words such as "the", "a", and "of" were likewise removed. Non-English stopwords were filtered out to augment the incapability of RapidMiner in handling insignificant data for clustering. The Filter Tokens by length stage deleted words with length less than 4 characters as minimum and 25 characters as maximum. The Stem Porter operator removed word affixes. For uniformity, the results of prior stages were transformed into lowercase formats. These were aggregated to come up with a corpus of pre-processed data.

3. Vector Creation

Vector Creation is a tool in RapidMiner that was used to calculate the frequency of words present in the corpus. This is specifically done using a method called Term Frequency – Inverse Document Frequency (TF-IDF). To expedite the processing time, percentual prune method was set with below percent value of 60.0, and above percent value of 100.0. This prunes the words that appear in less than 60% of the corpus.

4. Clustering

In order to identify the stand points of users in terms of term frequency weights and the themes that interest them, K-means technique has been applied. This technique is one of the simplest yet proven as one of the most reliable techniques in data mining. Generally it is described by a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). Below is its mathematical equation.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \qquad (1)$$

In other words, its objective is to find where $\boldsymbol{\mu}_i$ is the mean of points in S_i. Figure 2 illustrates as to how the K-Means algorithm clusters data.

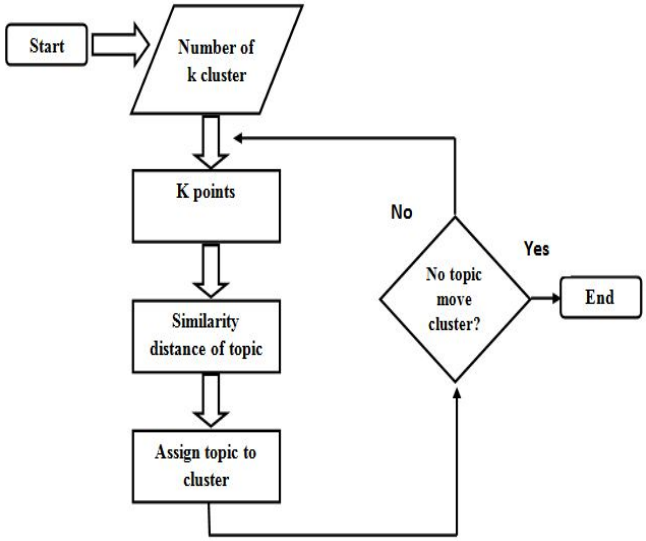

Figure 2. K-Means Algorithm

IV. Results and Analysis

After thoroughly processing the corpus of data by utilizing RapidMiner operators and methods, overriding and prevalent topics per cluster of users were identified. Results of processing the status and likes of facebook users are shown in the tables and figures below. The total occurrence of each topic from likes and status, the term frequency of each topic and the dominant topic from each cluster has been determined.

TABLE I. Facebook Status and Likes Topic Occurrences

Topic	Total Occurrences	Status and Likes Topic Occurrences per user	Likes	Status
food	13641	27	21	13620
game	7698	32	76	7622
cinema	6271	24	27	6244
pictur	5811	27	20	5791
movi	5675	31	22	5653
work	2617	28	15	2602
phone	2437	24	12	2425
sport	2160	24	29	2131
friend	1913	27	20	1893
video	1417	28	18	1399
song	1410	25	28	1382
love	955	30	25	930
school	762	28	69	693
philippin	643	27	195	448
citi	492	29	68	424
time	367	24	27	340
quezon	355	28	40	315
link	325	34	18	307
thank	181	33	19	162
head	132	33	66	66

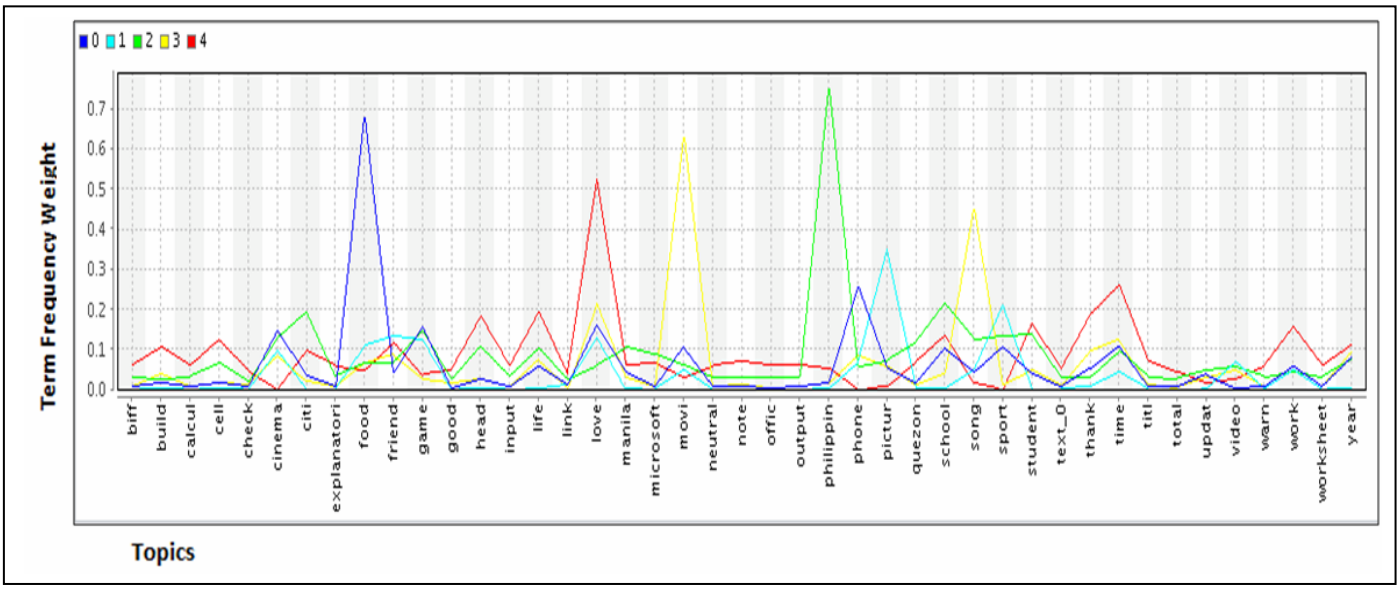

Figure 3. Clusters of Users Based on Term Frequency Weight

Dominant topics were identified based on their total occurrences. Table 1 shows the individual distribution of topics based on their occurrences for both likes and status. Values for each topic were used as bases in identifying the topics' frequency weights shown in Figure 3. Highlighted in Table 1 are the most liked topics. Top occurring topics include food, game, cinema, picture and movie.

TABLE II. DOMINANT TOPIC PER CLUSTER

Topic	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
biff	0.009	0.002	0.033	0.008	0.061
build	0.017	0.004	0.026	0.041	0.108
calcul	0.009	0.002	0.033	0.008	0.061
cell	0.020	0.003	0.066	0.017	0.126
check	0.010	0.002	0.020	0.015	0.047
cinema	0.149	0.097	0.131	0.087	0.000
citi	0.037	0.002	0.195	0.023	0.099
explanatori	0.009	0.002	0.033	0.008	0.061
food	**0.681**	0.110	0.069	0.068	0.048
friend	0.043	0.134	0.068	0.089	0.119
game	0.158	0.126	0.149	0.029	0.038
good	0.006	0.001	0.026	0.017	0.052
head	0.027	0.005	0.108	0.025	0.183
input	0.009	0.002	0.033	0.008	0.061
life	0.061	0.003	0.106	0.076	0.197
link	0.012	0.013	0.025	0.005	0.042
love	0.162	0.131	0.061	0.214	**0.527**
manila	0.047	0.003	0.106	0.030	0.064
microsoft	0.008	0.001	0.089	0.010	0.067

Topic	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
movi	0.107	0.049	0.064	**0.632**	0.030
neutral	0.009	0.002	0.033	0.008	0.061
note	0.009	0.002	0.033	0.011	0.072
office	0.006	0.001	0.033	0.005	0.062
output	0.009	0.002	0.033	0.008	0.061
philippin	0.019	0.003	**0.753**	0.018	0.056
phone	0.259	0.067	0.060	0.084	0.000
picture	0.057	**0.353**	0.073	0.060	0.007
quezon	0.019	0.002	0.119	0.013	0.072
school	0.105	0.003	0.217	0.042	0.137
song	0.045	0.052	0.126	0.451	0.017
sport	0.108	0.212	0.135	0.011	0.000
student	0.040	0.000	0.139	0.050	0.167
text_0	0.008	0.002	0.032	0.010	0.052
thank	0.054	0.009	0.033	0.097	0.188
time	0.107	0.047	0.098	0.125	0.261
titl	0.009	0.002	0.033	0.011	0.072
total	0.007	0.001	0.029	0.006	0.046
update	0.039	0.002	0.051	0.029	0.020
video	0.005	0.070	0.057	0.049	0.025
warn	0.009	0.002	0.033	0.008	0.061
work	0.061	0.051	0.044	0.061	0.158
worksheet	0.009	0.002	0.033	0.008	0.061
year	0.079	0.005	0.078	0.093	0.114

continuation of TABLE II.

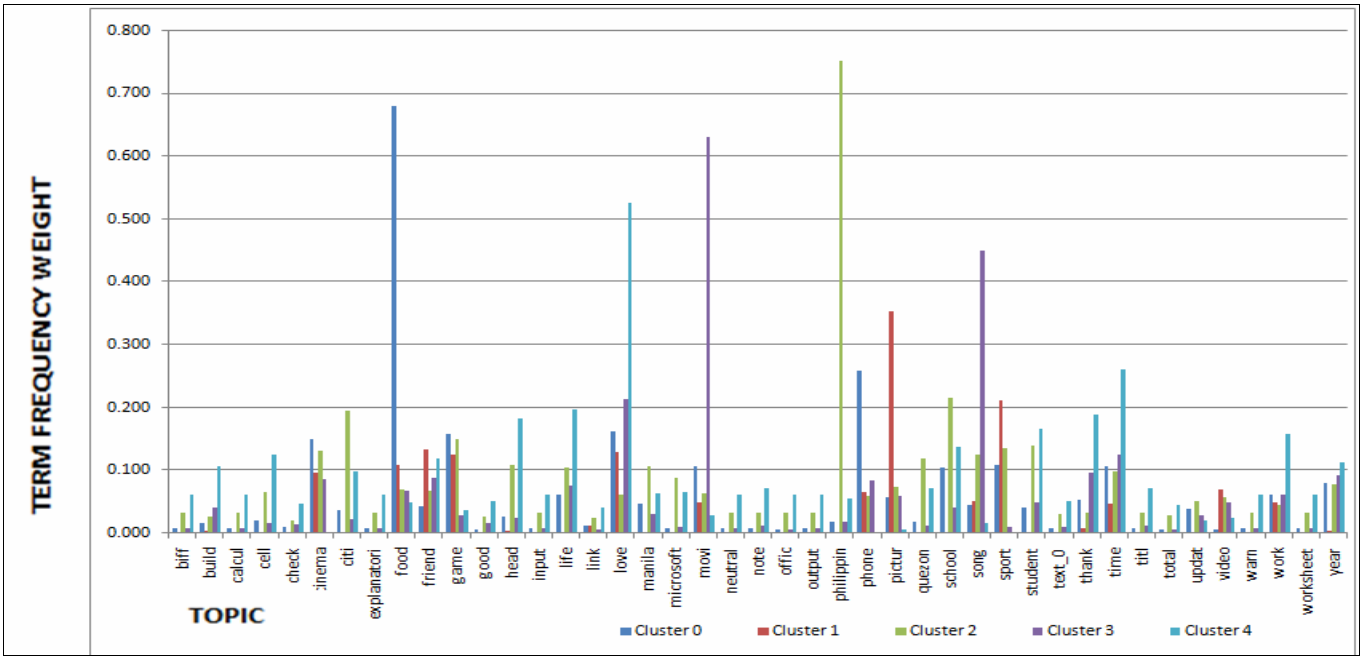

Figure 4. Histogram of Topics Based on Term Frequency

Disclosed in Table 2 are dominant topics per cluster. Food has gained the highest term frequency weight in Cluster 0, Picture for Cluster 1, Philippines for Cluster 2, Movie for Cluster 3, and Love for Cluster 4. These topics could serve as trending indicators of what possible business would most likely be embraced by specific cluster of users. Other prevalent topics are illustrated in Figure 4. Probable combinations of topics per cluster could likewise be drawn for outstretched network marketing.

V. CONCLUSION

In this paper, Faceboook users were clustered according to their explicit expressions in the form of status and likes. K-Means clustering technique was applied in order to identify the most dominant topics per cluster of users. After thoroughly processing the corpus of data by utilizing RapidMiner operators and methods, overriding and prevalent topics per cluster of users were identified. Results could provide insights for potential business and marketing campaigns that are pertinent to specific cluster of users.

ACKNOWLEDGMENT

This study was supported by the Commission on Higher Education (CHED) Faculty Development Program. Our sincerest gratitude to all facebook users who voluntarily participated in this study.

REFERENCES

[1] J. Surma and A. Furmanek, "Improving marketing response by data mining in social network", 2010

International Conference on Advances in Social Networks Analysis and Mining, pp 446 – 451

[2] N. Salamanos, E. Voudigari, T. Papageorgiou and M. Vazirgiannis, "Discovering Correlation between Communities and Likes in Facebook", 2012 IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing

[3] P. Limsaiprom and P. Tantatsanawong "Social Network Anomaly and Attack Patterns Analysis," 2010 6th International Conference on Networked Computing (INC), pp 1 – 6

[4] Dr. A. Fortino and A.Nayak, "An Architecture for Applying Social Networking to Business", 2010 IEEE

[5] K. C. Gull, S. C. G, A. B. Angadi and S. G. Kanakaraddi, "A Clustering Technique To Rise Up The Marketing Tactics By Looking Out The Key Users", 2014 IEEE International Advance Computing Conference (IACC)

[6] E. Pöyry, P. Parvinen and T. Malmivaara, "The Power of 'Like' – Interpreting Usage Behaviors in Company-Hosted Facebook Pages", 2013 46th Hawaii International Conference on System Sciences, pp. 1530-1605

[7] S.Lajmi, J. Stan, H.Hacid, E.Zsigmond, P.Maret, "Extended Social Tags:Identity Tags Meet Social Networks", 2009 International Conference on Computational Science and Engineering. Vol. 4 pp 181 – 187

[8] A. Friggeri, R. Lambiotte, M. Kosinskiand E. Fleury, "Psychological Aspects of Social Communities", 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

[9] Z. Jiang and C.Lu, "A Latent Semantic Analysis Based Method of Getting the Category Attribute of Words", 2009, pp. 141 – 146

[10] E. M.Mazla, S. Rahma, R. Ahmad and R.Kasbon, "Development of a Manufacturing Industry Success Rate Analyzer Using Data Mining Technique", 2010Information Technology (ITSim), International Symposium. Vol. 2 pp. 1032 – 1036

[11] C. Canali, S. Casolari, and R. Lancellotti"A quantitative methodology to identify relevantusers in social networks", 2010 IEEE

[12] M. Kosinskia, D. Stillwelland T. Graepel, "Private traits and attributes are predictable from digital records of human behavior", 2013 PNAS Early Edition

[13] X. Jin, C. Wang, J. Luo, X. Yu and J. Han, "LikeMiner: A System for Mining the Power of 'Like' in Social Media Networks", KDD'11, August 21–24, 2011

[14] H. Wu, K. Liu and C. Trappey, "Understanding Customers Using Facebook Pages: Data Mining Users Feedback Using Text Analysis", Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design

[15] T. Verma, Renu and Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014 – www.ijais.org

[16] S. Bhatia, J. Li, W. Peng, and T. Sun, "Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems", 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[17] J. Kim, D. Choi, B. Ko, E. Lee, and P. Kim, "Extracting User Interests on Facebook", Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2014, Article ID 146967, 5 pages http://dx.doi.org/10.1155/2014/146967

AUTHORS PROFILE

Paula Jean M. Castro graduated with a BS Information Technology degree from Technological Institute of the Philippines-Quezon City (TIP-QC) in 2012. She became a full-time faculty member of TIP-QC College of Information Technology Education in April 2012. Pursuing her degree in Master in Information Technology, she is currently undertaking her Thesis on Data Mining.

Rosmina Joy M. Cabauatan is Assistant Professor of both Graduate School and College of Information Technology Education of Technological Institute of the Philippines. With her research interests in educational games, and applications of data mining in the computing and educational disciplines, she has been commended for her contributions to institutional researches. Pursuing her degree in Doctor in Information Technology, she is currently undertaking her Dissertation on application of data mining in mobile computing.

Bartolome T. Tanguilig III took his Bachelor of Science in Computer Engineering at Pamantasan ng Lungsod ng Maynila, Philippines in 1991. He finished his Master's Degree in Computer Science from De La Salle University, Manila, Philippines in 1999, and his Doctor of Philosophy in Technology Management from Technological University of the Philippines, Manila in 2003. He is currently the Assistant Vice President for Academic Affairs (AVPAA) and concurrent Dean of the College of Information Technology Education and Graduate Programs of the Technological Institute of the Philippines, Quezon City.

Dr. Tanguilig III is a member of the Commission on Higher Education (CHED) Technical Panel for IT Education (TPITE), the chair of the CHED Technical Committee for IT (TCIT), the founder of Junior Philippine ITE Researchers (JUPITER), board member of the Philippine Society of IT Educators (PSITE), member of the PCS Information and Computing Accreditation Board (PICAB), member of the Computing Society of the Philippines (CSP), and a program evaluator/accreditor of the Philippine Association of College and Universities Commission on Accreditation (PACUCOA).