２０１６年度　修士論文

Characterizing Relationships Between Public Posting Activities, Interest groups, and Profiles of SNS Users

指導教員　岩井原 瑞穂 教授

早稲田大学大学院情報生産システム研究科
情報生産システム工学専攻　データ工学研究

44151005－1　　Dewi Nurfitri Oktaviani

Abstract

In this research, we want to characterize the relationships between public posting activities on particular interest groups through a data-driven approach. Our objective here is to discover user types from their sets of profile values and posting behaviors, which can be collected from their public timeline posts and profile pages on Facebook. The user types are derived from clustering analysis with a data-driven approach, which enables us to classify users who post frequently on public pages, find out the new things in online communication activities that happened in social media.

We adopt two-level clustering. First, we cluster the posts into groups of posts based on their metadata information, including subjectivity and polarity scores, length of posts and number of interactions to reveal users' posting behavior. Second, we cluster the users into groups of users based on the posting features as well as activity features. Posting style and activity features are normalized and combined as a feature vector. We apply k-means clustering to normalized 11-dimension feature vectors to identify characteristic user types based on their posting styles, sentiment, and influentiality. We discovered five distinct user types that characterize public posting activities as follows: (1) cynical-active users; (2) thoughtful-influential users; (3) joyful-influential users; (4) cynical-lone users; and (5) joyful-lone users. We observe that these user types present distinct distributions between user groups of different interests.

Keyword: SNS, K-means clustering, user types

Table of Contents

List of Figures

List of Tables

1 Introduction

In recent years, social network services (SNSs) grow rapidly and attract a large number of users in the world. SNSs allow users to create a public profile and provide space for users to express their opinion, interact with others, share contents, and upload photos/videos of recent activities. Activities of SNS users are often triggered by their interests. For example, users show their concerns and attentions to their visited places or purchased products by posting reviews in their SNS accounts. Also, users who like pets frequently upload their pets' photos, etc. Such activities SNS users are potential dataset to reveal the relationship between posting activities and interest of users, which are invaluable resources for advertisement and monetization of SNSs.

In this research, we would like to characterize public posting activities and reveal how they are related to users' interests, through a data-driven approach. Thus, firstly, we define user type based on their sets of behaviors that commonly co-occur through their online record activities on Facebook. Secondly, to define characteristics of interest groups, we examine how the user types are distributed along the interest groups. An interest group is a place for a number of users to interact, communicate and express opinions about the common interests.

We collected public posts of users, extracted their metadata information, such as length of the posts, the frequency of interactions, and we applied sentiment analysis on the post texts to evaluate subjectivity and objectivity scores. Then we apply an unsupervised clustering method on the post texts, to discover typical user types, described by posting styles, activity-related features, and user profiles.

Our objective here is to discover user types from their sets of activities and posting behaviors, which can be collected from their public timeline posts and profile pages on Facebook. The user types are derived from clustering analysis with a data-driven approach, which enables us to classify users who post frequently on public pages.

Our user sets are sampled from a number of interest groups, where each interest group is a place for member users to interact, communicate and express opinions about common interests. However, post texts we sampled are from users own public

timelines, so postings are not directed to group members, but to the public. We regard that a user's participation to an interest group implies that the user has an interest of the subject of the group. We study whether significant differences exist between users who are joining to various interest groups. By comparing distributions of user types, we try to examine whether an interest group is dominated by a particular user type and to find similarities of posting behaviors and moods in members of interest groups. Our feature set captures emotional tendencies of users' postings, through sentiment analysis as well as posting frequencies. On the other hand, our feature set is independent from the semantics of interests, such as keywords for interests, which enables us to compare user types over users of semantically unrelated interests, such as pet lovers and political enthusiasts.

We discovered the five distinct user types as follows: (1) The cynical-active users represent the group of users who actively writes posts and shares news from others but seldom receives feedback; (2) The thoughtful-influential users represent the group of quite popular and interactive users who often writes formal and serious posts that mostly earn feedback; (3) The joyful-influential users represent the group of users who often makes friends, writes positive posts that frequently gain feedback; (4) The cynical-lone users represent the group of users who has the least number of activities on Facebook, once they write posts, they convey negative sense that maintain low interaction; and (5) The joyful-lone users represent those who often deliver the positive sense of posts but rarely gain feedback from others. We observe that these user types present distinct distributions between different interests.

Revealing relationships between interests of users and their posting activities and profiles is useful for various applications in SNSs, such as (1) recommendation system that exploit potential types of users. For example, SNSs should provide more portion of friend suggestions on newsfeed for those who likely to make friends; (2) an effective moderation assistance for group administrators is possible, by classifying users who request to join a group into our user types; and (3) advertisement campaigns for online marketers can be personalized by our user types.

The rest of this thesis is organized as follows. Chapter 2 explains the current research on SNS user analysis. Chapter 3 describes our methods to collect and extract dataset from SNSs. Chapter 4 explains about our experiments, including feature extraction, preprocessing, clustering and validation. Chapter 5 presents our evaluation results of clustering and identified user types. Finally, Chapter 6 presents a conclusion and future work.

2 Related work

Behavior of users in SNSs is a significant topic of SNS researches, producing a large volume of publications, which are exploring and revealing characteristics of user behaviors in SNSs. One example is to discover features for predicting user attributes, such as user income prediction [13], the dark triad personality prediction [12], age and gender prediction [9], latent user properties prediction, such as demographic attributes [14], and five users' posting purposes on using SNSs, including posting group or fellows' photos, propagating positive social news, posting self-photos, propagating positive quotes, and posting personal details [10].

Various researches have been done on modeling user interests in SNSs. Most of studies are about inferring interests of SNS users, from daily posts of users using techniques such as topic models. However, daily posts often contain conversations about daily activities of users, which are difficult to identify the interest of users. Bhattacharya et al., [5] discovered topical expertise of Twitter users, by collecting users' lists, then extracting the most common terms that appear in the list names and list descriptions. A user is said as an expert on a topic if and only if the topic appears at least 10 times in the metadata of lists containing him. Their results showed that their methodology is far superior than topic extraction from the contents of tweets by LDA. Kim et al. [7] took into account the number of likes in Facebook and topic contents, to predict user interests.

In contrast to the existing researches, we aim to characterize and typifying user activities based on their publicly-posted SNS contents. We define user types based on

their posting behaviors and profile attributes that are visible from their online activities on Facebook. Our user types are derived through clustering with a data-driven approach, where posting behaviors are abstracted from sentiment and frequency-oriented metrics, that are independent from topical keywords. Therefore, our approach can capture similarities on user moods and behaviors over different domains of interests, so that we can compare user groups of various interests, distributions of the user types.

Many works have been done on characterizing user in SNSs, such as: a research conducted by [11] identified and compared viral characteristics across the mixed media types (text, photos, videos) found on Facebook. Previous research conducted by [8] clustered user types from sentiment distributions of users' public posts, through three level polarity, five level polarity, and five level sentiment to aggregate the post-level scores to user-level vectors. The limit threshold for each level were manually determined.

There are several improvements compared to previous research [8]: (1) we do not determine the threshold manually, but let a clustering algorithm to determine the flexible threshold; (2) we utilize the post metadata information, such as: length of posts and number of interactions (comments, likes and shares), instead of only subjectivity and polarity scores, to reveal the users' posting styles; and (3) we cluster the users into the types of users based on the posting style features as well as activity features.

3 Data Collection

In our research, the experimental data are collected from Facebook, which is currently one of the most popular SNSs providers. We sampled user sets by collecting Facebook User IDs through particular Facebook groups, crawling their public profiles and a set of public posts from a certain range of time. We used a dataset from our previous research's datasets [10] [8] and added a new user set.

There are users collected from previous research, including users from Business, Politics, Pets, and Music as presented on Table 1. There are 207 users from several

public Facebook groups of different interests, such as Business, Politics, Pets, Music and a random group from the previous dataset. We obtained 42 users from the music group, 32 users from the pet group, 63 users from the political group, and 70 users from the Trade group.

Table 1 Number of users from various interest group

No.	Interest Group	Number of Users
2	Music	42
3	Pets	32
4	Politics	63
6	Trade	70
Total		**207**

3.1 Selecting users through Facebook Graph API

Our new datasets are collected from three different Facebook groups, consisting of Food, Community, and Travelling. In the first step, we need to perform user selection. We selected active users from public groups under Food, Travelling and Community group on Facebook through Graph API. We considered users who just wrote post/discussion on the group walls as the active users.

Facebook Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). However, certain API call operations require permission from users. Using this interface, we accessed to active members of the public Facebook groups which are returned as JSON objects. As we can see in Table 2, we were able to obtain 479 users IDs from three different groups during our research. We gathered 75 users from Food group, 310 users from Traveling group, and 96 users from Community group.

Table 2 Number of users from new dataset

No.	Interest Group	Number of Users
1	Food	75
2	Community	94
3	Travelling	310
Total		**479**

3.2 Web-crawling on Publicly Available User's Content

For the purpose of our study, we collected publicly available profile information and posting dataset from Facebook. Since we have collected the list of active members of the public Facebook groups, we just need to develop a crawler that reads specified single active user IDs.

We employed bots or spiders that read the HTML documents of each user ID to provide graphical interfaces to users via browser automation. We used the Python and Selenium framework [3] to automate the Chrome browser to open a page, log in to Facebook, access the user's timeline based on the user ID, scroll and save the page until the posts reach the year of 2014.

Since we collect only public posts, we do not need permission from users to access their data. However, we keep users' identities anonymous in this thesis.

3.3 Extracting Relevant Information from Users

The stored web pages generated by the automated web browser are parsed and extracted with Beautifulsoup4 [1] into text files. We automatically extracted users' posts and the following six observable activity features from users as reflected in Table 3.

Table 3 Activity features description

Activity Features	**Description**
Number of friends	indicate the size of the social graph the user has. A more number of friends indicates a more influence the user has
Number of self-uploaded photos	imply whether the user likes to share his/her activity through photos in SNS
Number of self-uploaded videos	imply whether the user likes to share his/her activity through videos in SNS
Number of profile photos	suggest if the user is eager to disclose his/her face.
Number of shared news	indicate whether the user likes to share text contents from others or write his/her own posts.
Number of posts	imply if the user likes to write his/her own opinion in the SNS.

3.3.1 Computing Sentiment Score from User's Posts

Sentiment analysis is performed to each user's text post, using the sentiment analysis tool named TextBlob [4]. TextBlob is a Python library to calculate sentiment score of

text, producing a polarity and a subjectivity score for an English text. The polarity score is a float within the range -1 to 1. The subjectivity is a float within the range 0 to 1.

Polarity score can tell us how the expressed opinion of the text is inclined toward positive, negative or neutral. Subjectivity score can tell us the degree of how the text is objective or subjective. Sentiment can be evaluated based on the presence of affect words such as happy, sad, afraid, and bored.

We mainly omitted non-English posts from the sentiment analysis using the Natural Language Toolkit (NLTK) library [2]. After filtering non-English posts, as can be seen in Table 4 we obtained 36,912 English posts from the previous dataset and 25,753 English posts from the new dataset. In total, we obtained 62,665 English posts.

Table 4 Number of collected posts from users

Type of Posts	**Number of posts**
Posts from Music, Pets, Politics, and Trade groups	36,912
Posts from Food, Community and Travelling groups	25,753
Total	**62,665**

3.3.2 Extracting Users' Activity Features

Users' activity features are numerical features on how much degree a user is active on a particular aspect. Table 5 shows the description, count, mean, minimum, maximum and standard deviation values from the activity features.

Table 5 Summary of activity features

Features	**Type**	**Mean**	**SD**	**Min**	**25%**	**50%**	**75%**	**Max**
Number of Friend	int64	611.1622	1108.023	0	0	0	735.75	5,000
Number of Profile Photo	int64	48.92553	76.58982	0	6	22.5	62.5	1,146
Number of Post	int64	182.4827	177.711	2	55	115	233	910
Number of Shared News	int64	75.47075	83.51243	0	17	46.5	103.25	478
Number of Self-Upload Photo	int64	290.2766	999.0396	0	6	35	207.5	17,547
Number of Self-Upload Video	int64	0.212766	2.303808	0	0	0	0	60

Considering that not all users disclose certain features, such as their friend lists, the missing data is replaced with zero value. The zero value depicts either the users hide the activity values from public or the users do not post those activities.

4 Proposed Methods

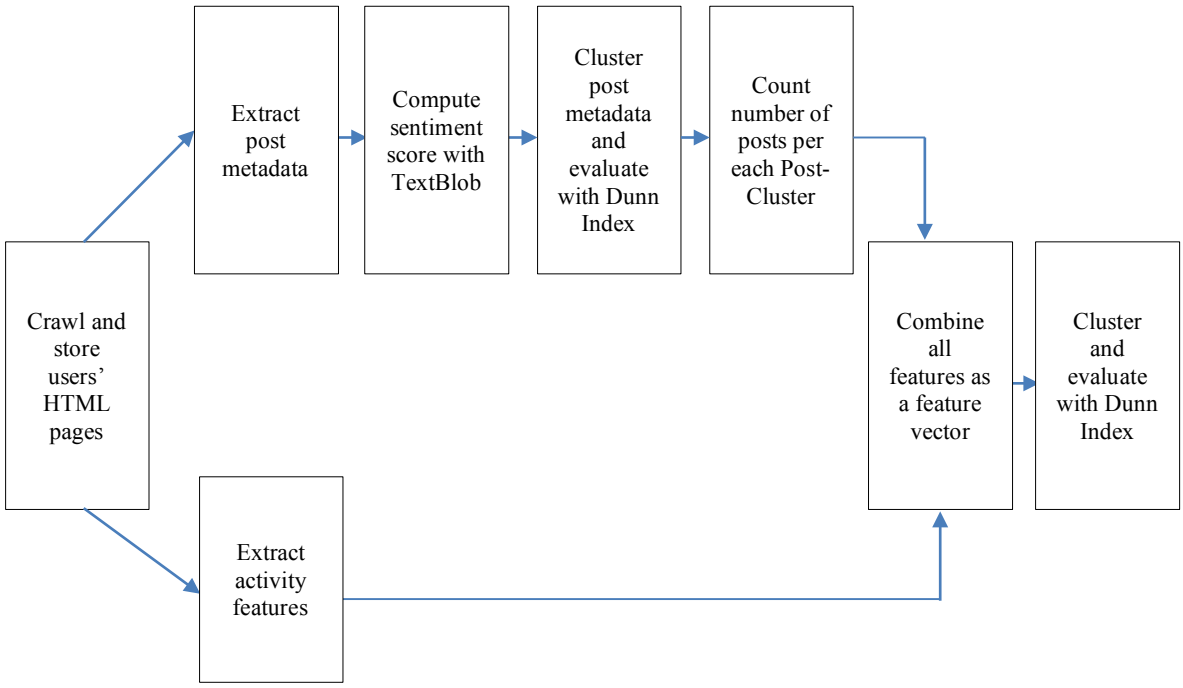


Figure 1 Our proposed method

As explained in the Chapter 3, we collected only publicly available profile information and posting dataset from Facebook. We develop crawler that read every single active user ID, store and extract the HTML pages.

There are two level of extracted data namely, post-level and user-level data. Post-level data is defined as data that describes the individual post contributions submitted by the users. The user-level feature is a feature defined as either the aggregation of post-level data or a single data that describes users' characteristic in the particular activities. We extract users' posts and following six observable activity features from users:

number of posts (text), number of shared news, number of self-uploaded videos, number of profile photos, number of self-uploaded photos, and number of friends.

As can be seen from Figure 1, to characterize users, we possess two level of clustering. First, we cluster the posts into the group of posts based on its metadata information, including subjectivity and polarity scores to reveal the sense of posts, length of posts and number of interactions. Second, we cluster the users into the group of users based on the post-cluster features as well as activity features. Posting style and activity features are normalized and combined as a feature vector. Thus, we run K-means cluster on a normalized feature vectors contained 11 data points to identify several user types based on their similarity activities.

We interpret the characteristic of each cluster once the clusters are generated and give a particular label to define the clusters, namely user types. Then, we examine how the characteristic of each Interest Group based on user types distribution.

5 Experiments and Evaluations

5.1 Experiments

5.1.1 Post-Level to User-Level Feature through Clustering

As part of our analysis, we want to discover the latent characteristics of posts that implied from the post metadata information. Post metadata information is the additional data that provide information about each single post. With particular post metadata, we can express supporting information from the posts in a way that doesn't detract the content itself. Some of the relevant post metadata are the length of the posts, the mood of the posts that is expressed from the subjectivity and polarity score of the posts, the number of interactions, namely the number of received comments/likes and the number of people who shared the post.

Note that not every user write posts in English, a further pre-processing step is necessary to filter our dataset from non-English posts. We aggregated the posts into a

user-level through clustering. We clustered the posts based on its metadata information. Table 2 shows the description of the posts' metadata information. As we can see from Table 6, users rarely give comments, they tend to share or give a thumb to posts. Sharing and liking post seem easier to do than giving comments. Due to high variability, we convert features into their natural logarithmic forms.

Table 6 Description of the posts' metadata information

Post Metadata Features	Type	Mean	SD	Min	50%	Max
Length of Post	int64	141	225	1	73	2,978
Number of Comments	int64	3	7	0	1	326
Number of Likes	int64	5,517	1,196,831	0	7	296,296,300
Number of Shares	int64	1	12	0	0	871

We aggregate the posts into a group of posts (user-level features) by applying K-means on their metadata information. We run a K-means on a normalized post feature vector contained 6 data points within k-value ranges from 2 to 10. The optimum value of k is computed based on the Dunn-index evaluation. After the clusters are generated, we compute the number of post for each cluster. These generated clusters namely, PostCluster0, PostCluster1, PostCluster2, PostCluster3, PostCluster4 and PostCluster5, are considered as features that describe the posting styles expressed by users. The clustering result is described in Section 5.2.1.

5.1.2 Feature Pre-processing

For the preprocessing, we transformed the values of behavior features using log function. The log function reduces the large raw values of behavior features into a smaller value and makes skewed distributions more symmetric.

Feature scaling matters for some algorithms, such as K-means with a Euclidean distance measure if want all features to contribute equally. Therefore, before clustering the posts, we normalize the features with the Equation 1.

$$x_{new} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

5.1.3 Clustering User

In this research, we aim to define user type based on their sets of behaviors that commonly co-occur through their online record activities on Facebook. The user types are defined through clusters with a data-driven approach, thus, we can find out the new things in online communication activities that happened in social media. Consequently, this approach can be enhanced with user-level features. User-level features are relevant to this because they can fully describe a user.

Table 7 Description of user-level features

User-Level Features	Type	Mean	SD	Min	50%	Max
Friends (f_1)	int64	632.6472	1137.771	0	0	5000
NoProfilePhotos (f_2)	int64	47.09038	75.48116	0	21	1146
UploadVideoSum (f_3)	int64	0.139942	0.764701	0	0	14
NoPosts (f_4)	int64	175.0889	171.3218	8	112	910
SharedNewsSum (f_5)	int64	75.98688	82.56235	0	48.5	478
UploadPhotoSum (f_6)	int64	316.2668	1042.327	0	47.5	17547
PostsCluster0_ratio (f_7)	float64	0.015129	0.059074	0	0	0.69697
PostsCluster1_ratio (f_8)	float64	0.48461	0.197257	0	0.482673	1
PostsCluster2_ratio (f_9)	float64	0.017932	0.04618	0	0	0.470588
PostsCluster3_ratio (f_{10})	float64	0.184255	0.181841	0	0.136148	1
PostsCluster4_ratio (f_{11})	float64	0.298074	0.165051	0	0.285714	1

We combine all the features as shown in Table 7, compute it as a feature vector $[f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}]$ as input for the K-means clustering algorithm to identify several user types based on their common/similarity behavior. Clustering is a task to partition datasets into subsets where in-class members are "similar" in some sense and whose cross-class members are "dissimilar" [15]. K-means clustering is the most widely used clustering algorithms. K-means compute the distance between features with uses Euclidean distance as given by Equation 2.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (2)$$

Therefore, K-means clustering is sensitive to the magnitudes of the variables, especially by outliers. Feature scaling matters to measure all features to contribute equally. We applied the K-means clustering to identify clusters of these feature vectors, and thus divide our Facebook users into groups of similar behavior. Since these features

have different units, we first normalized the feature values so that each normalized feature ranged from 0 to 1. The common characteristics of each cluster are interpreted through the centroid value of each feature which is explained more detail in Section 5.2.2.

5.1.4 Dunn-Index Validation

We evaluate the quality of clustering results based on the internal evaluation, called Dunn index. These cluster validity indices have been introduced in the paper [6]. The index definition is given by Equation 3. Dunn-Index value derives an optimal number of k by comparing the inter-cluster distance with intra-cluster distance.

$$DunnIndex(C) = min_{1\le i\le N}\left\{min_{i+1\le j\le N}\left(\frac{dist(c_i, c_j)}{max_{1\le i\le N}(diam(c_i))}\right)\right\} \quad (3)$$

Inter-cluster distance calculates the distance between the point and the centroid of other clusters. Intra-cluster distance calculates the maximum distance between the point and the centroid. In a simple way, Dunn Index finds the comparison between minimum inter-cluster distance and maximum intra-cluster distance. The higher Dunn Index value, the better k- value.

5.2 Evaluation and Result

In this Section, we will explain: (1) the latent roles in aggregation result from post-level to user-level features; (2) the user types derived from User Type on posting styles and activities; and (3) the distribution of defined user types in different interests group.

5.2.1 Common posting styles through clustering on post metadata information

We aggregate the posts into a group of posts (user-level features) by applying K-means on their metadata information. Figure 2 shows the optimum value of K based on the Dunn-index evaluation.

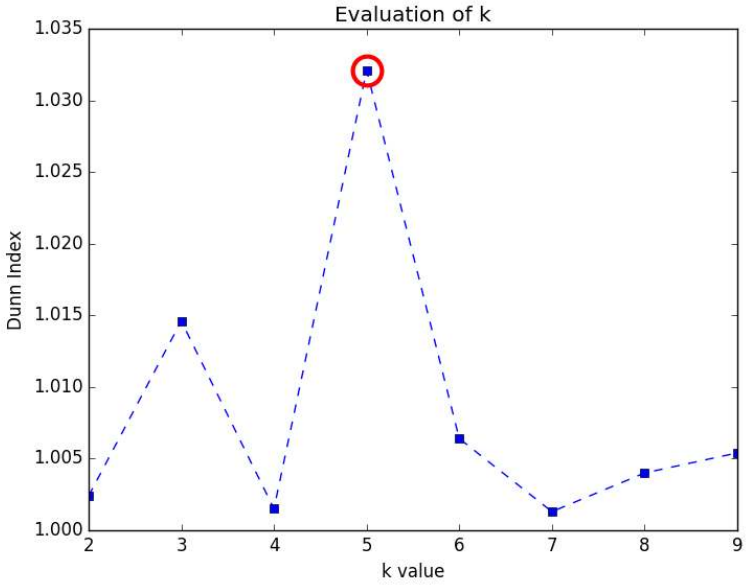


Figure 2 K-value comparison on post-clustering

Table 8 shows five post-clusters which have unique characteristics marked by the length, the subjectivity, and polarity of the post, the number of received comments, likes and people who shared. Since the clustering is done on normalized scale, we obtained the normalized centroid of features. In order to interpret the value, we reverse scale the centroid and interpret the characteristic of each cluster is the real value of centroid.

From the post cluster interpretation, we can conclude several common posting styles such as the negative posts receive less feedback from others, on the contrary, the neutral posts always receive feedback from others, either comments or likes. The length of posts seems do not deal with the number of earned feedback.

Table 8 Post-Cluster interpretation result

PostsCluster	**Interpretation**
0	Long-length, objective, and neutral posts that are mostly shared by others receive likes and comments
1	Short-length, objective, and negative posts that are almost never shared by others, seldom receive likes and almost never receive comments
2	Very long-length, subjective, neutral posts that are never shared by other, but sometimes receive likes and comments
3	Short-length, objective, neutral posts that are almost never shared by other, but mostly receive likes and comments
4	Short-length, subjective, positive posts that are never shared by other, never receive likes, but sometimes receive comments

The posting styles features are generated from the number of post for each PostCluster.

5.2.2 User types derived from User's Posting Activities and Profiles

We run a K-means on a normalized user-level feature vector contained 11 data points within k value ranges from 2 to 10. The optimum k value is computed based on the Dunn-index evaluation.

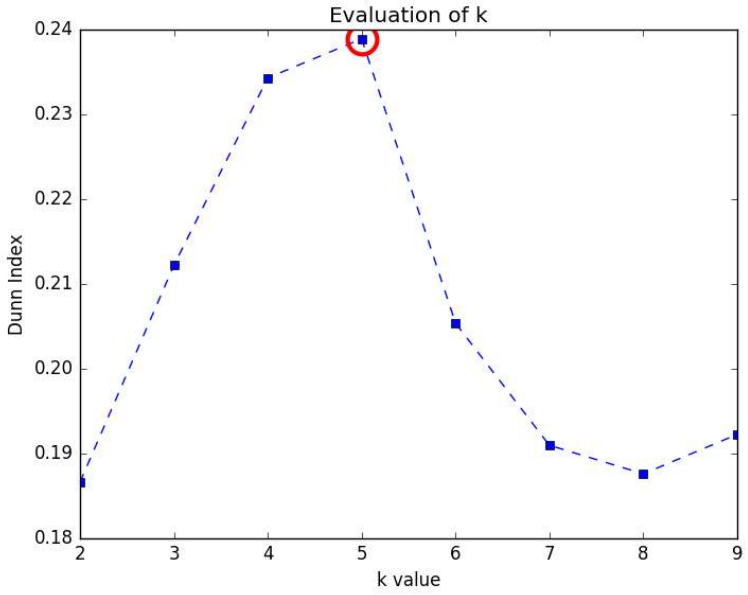


Figure 3 K-value comparison for user-clustering

As we can see from the Figure 3, the optimal value for k is 5. Therefore, we run K-means with five as the k-value. In the next explanation, the term of clusters of users are replaced with user types. We visualized the separated clusters of users in 3D plots as shown in Figure 4 and Figure 5. As we plotted in 3D graph, we reduce the non-significant features and choose the most important features that distinguish the user types. These important features are selected from the standard deviation through clusters. We notice the number of friends, the number of posts and PostCluster3 are the distinguishing features among user types.

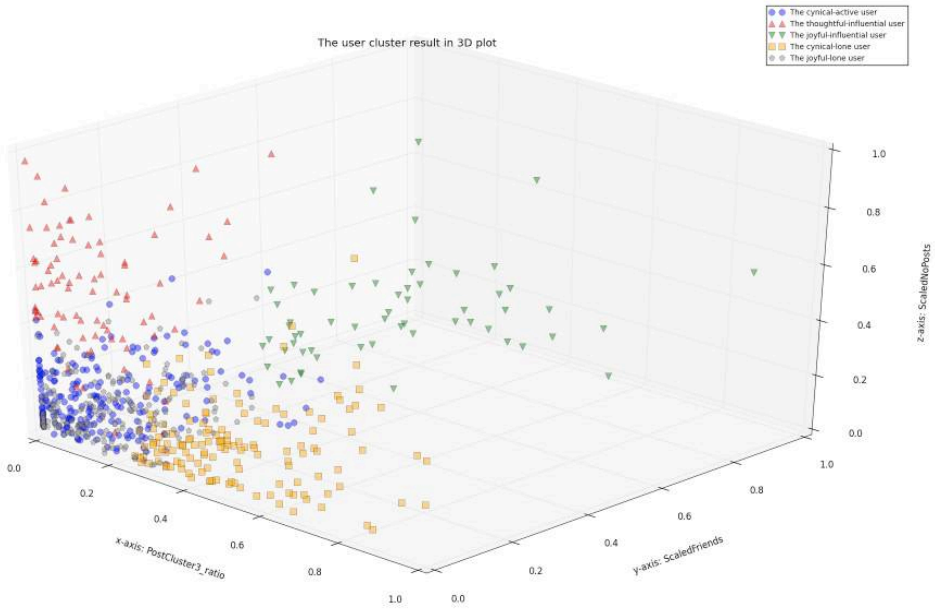


Figure 4 Types of Users in 3D Visualization (xyz-view)

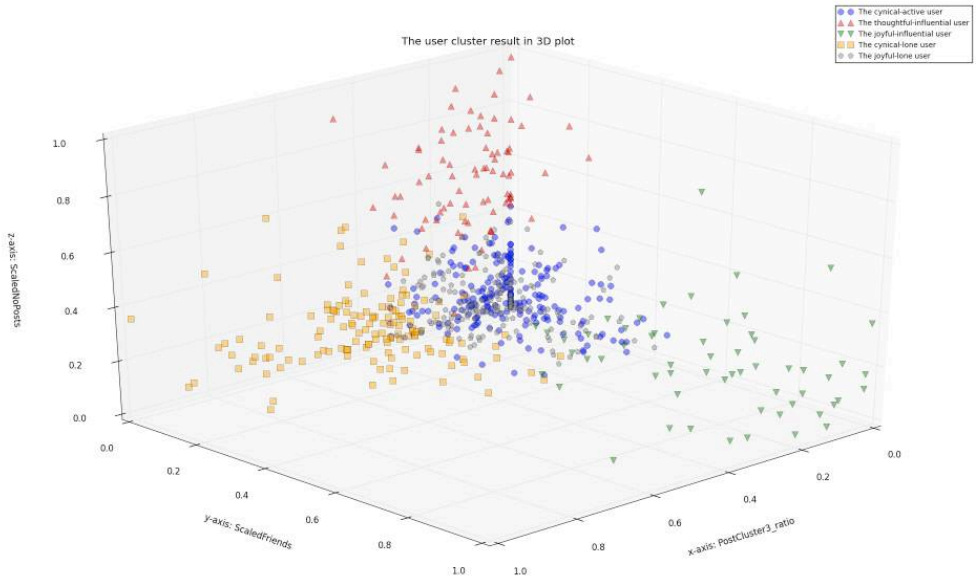


Figure 5 Types of Users in 3D Visualization (zyx-view)

As we can see from the 3D plot, the five types of users are well separated. 'the joyful-influential users' occupy the high value area of number of friends (y-axis), while the high value of x-axis (ratio of PostCluster3) is saturated by 'the thoughtful-influential users'. 'the cynical-active users' are scattered on the middle value of the z-axis (number of posts).

To be more detail, the Figure 6 shows diagram of how all features varied for each user type in a normalized scale of the average value of features. These user types are separating users based on several characteristics.

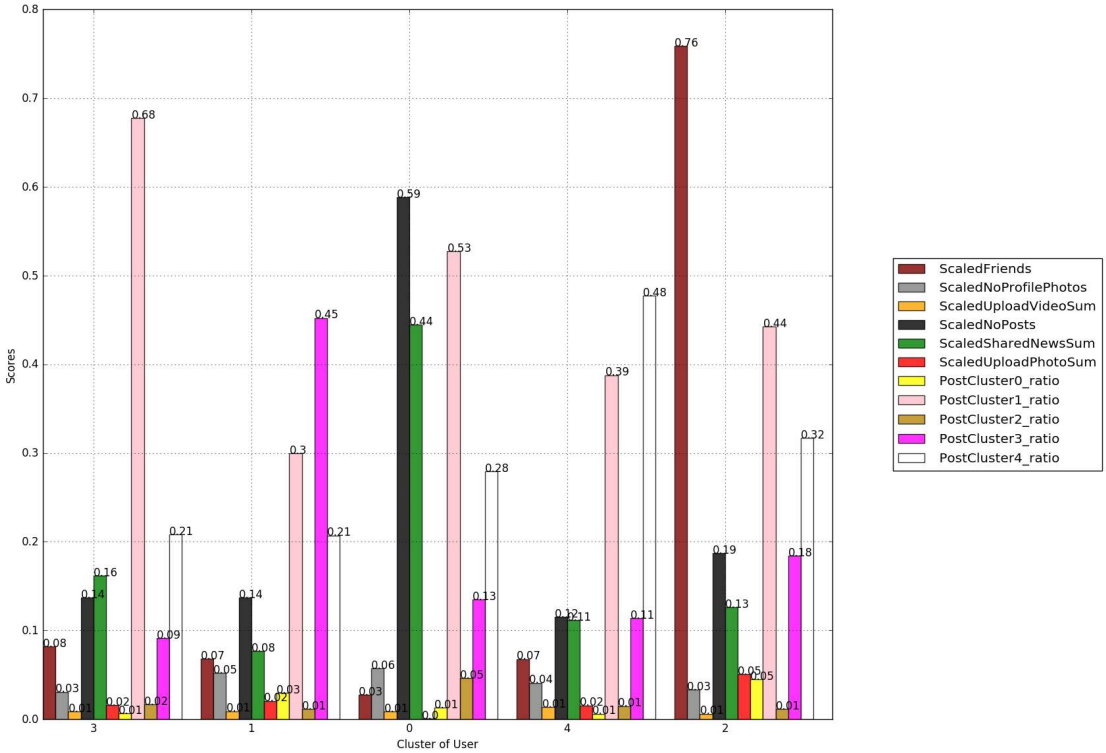


Figure 6 Characteristic of clusters of users in a clustered bar

We point out that the mean value of several features (Self-Uploaded Videos, Number of Profile Photos, the relative value of PostsCluster2) do not vary much across clusters. The overall distribution of these three features shows a narrow range of values and most users have a similar value for each of the three features. Despite its user type, every user has more than 30% PostsCluster1. PostsCluster1 is a group for posts that is written in short-length, convey the objective and negative sense and rarely gain feedback from others.

Table 9 presents a summary of characteristics between different clusters through the dominant posting style features and activity features where dominant means the features' value are above its average.

Table 9 Dominant Features on Clusters

User Type	Posting Style Features	Activity Features
User Type 0	PostsCluster1 and PostsCluster2	Number of posts (text), shared news, and profile photos
User Type 1	PostsCluster3 and PostsCluster0	Number of profile photos
User Type 2	PostsCluster4 and PostsCluster0	Number of friends and self-uploaded photos
User Type 3	PostsCluster1	
User Type 4	PostsCluster4	Number of self-uploaded videos

In order to easily interpret the user cluster, here we define several terms based on the frequency activity features and posting style features. We divided the features into two different characteristics, posting style and activity features. We noticed that there are several common characteristics on users types as presented in Table 10.

Table 10 Labeling based on Posting Styles Features

Features	Labels	Description
Posting Styles	Cynical	For users who write posts with objective and negative sense, either written in long or short text
	Thoughtful	For users who write posts with objective and neutral sense, either written in long or short text
	Joyful	For users who write posts with subjective and positive sense, either written in long or short text
	Influential	For users who often make interactions with others through likes and comments.
	Lone	For users who write posts that gain less feedback from others, such as: seldom receive likes and comments.
Activities	Active	As several features, such as Self-Uploaded Videos, Number of Profile Photos do not vary much across clusters, we decided to use the number of Friends, postings and sharing posts as an indicator. For users who have high activities in Facebook that are indicated by the more number of friends, postings sharing posts and profile photos.
	In-active	For users who have low activities on Facebook that are indicated by the less number of friends, postings sharing posts and profile photos.

Need to be noted that the user type label for each cluster is given from the most dominant posting styles and activity features. We summarized the five distinct user types that define online activities happened on Facebook as follows:

5.2.2.1 User Type 0: The cynical-active users

This cluster represents the group of users who actively writes posts and shares news from others, but has the least number of friends and least number of self-uploaded photos. The higher number of posts and shared news indicate the higher activities in Facebook.

Looking at their posting styles, they have high ratio value of PostsCluster1 which means they often write short-length posts that deliver objective-negative sense that almost never shared by others, seldom receive likes and almost never receive comments, as shown in Figure 7. Though it is not stand out, this cluster has high ratio number of PostsCluster2 (very long-length, subjective, neutral posts that are never shared by other, and sometimes receive likes and comments). Here are examples of their post:

- Another one bites the dust...
- Wake Up! Obama's destruction of America plan is working!

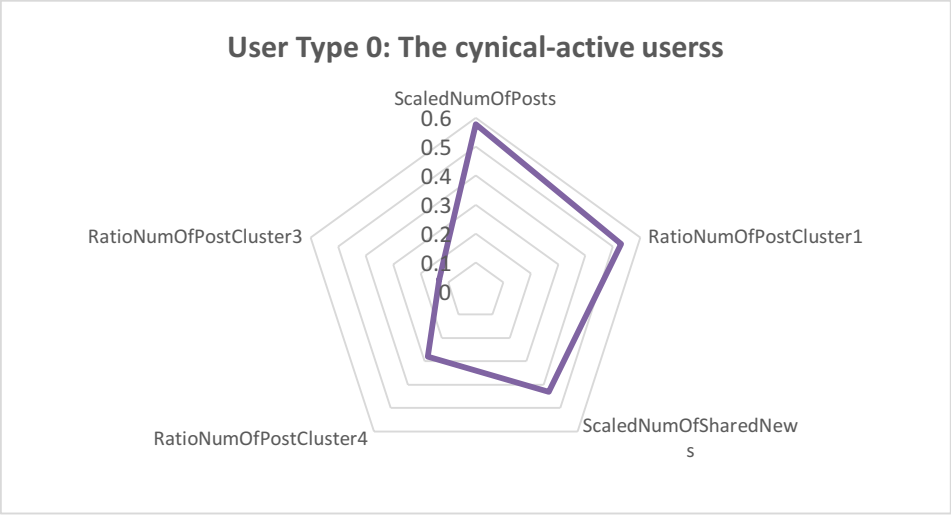


Figure 7 Top-5 most prominent characteristic for 'The cynical-active users'

5.2.2.2 User Type 1: The thoughtful-influential users

As we can see from Figure 8, this cluster represents the group of users who has a high ratio of PostsCluster3, which means they often write short-length posts that convey objective and neutral sense and almost never shared by other but mostly receive likes

and comments. They are quite popular and interactive as they often write posts with the formal and serious topic but mostly earn feedback from others. Here are examples of their post:

- Jeunesse has great products of the future that you don't want to miss out on. Please take two short minutes that could change your life and learn about Finiti and what it does to help your body. http://www.stayinforeveryoung.jeunesseglobal.com
- Obama has done everything backwards, what kind of brain is running our country

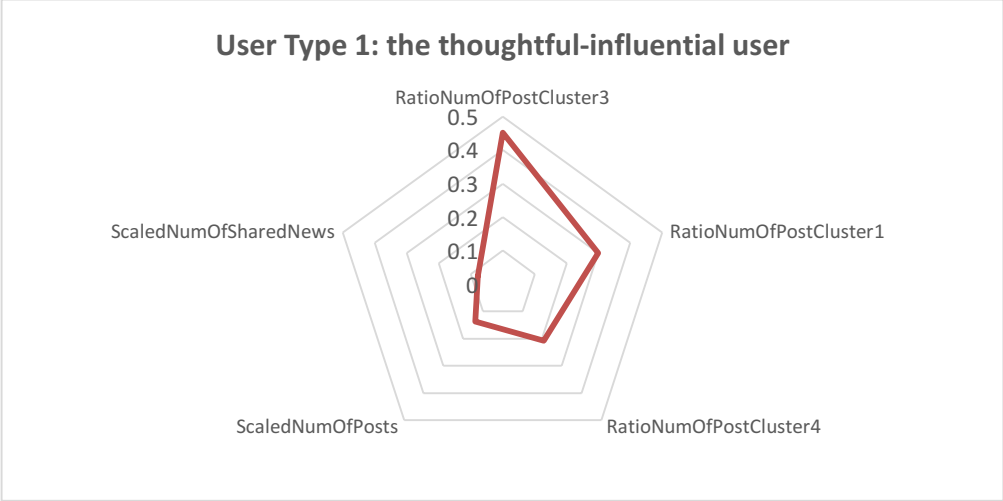


Figure 8 Top-5 most prominent characteristic for 'The thoughtful-influential users'

5.2.2.3 User Type 2: The joyful-influential users

The group of users who often makes friends as they have a high number of friends and many self-uploaded photos in average as shown in Figure 9. Compared to other users, they have high ratio of PostsCluster4 (short-length, subjective, positive posts that are never shared by other, never receive likes, but sometimes receive comments), and have high ratio of PostsCluster0 (long-length posts with objective and neutral sense that are mostly shared by others, receive likes and comments). Here are examples of their post:

- Love it!
- So true! I think I should take this seriously from now.

From the view of activity features, it seems the number of friends is correlated to the number of self-uploaded photos. As a comparison, 'the cynical-active users' have least number of friends and least number of self-uploaded photos.

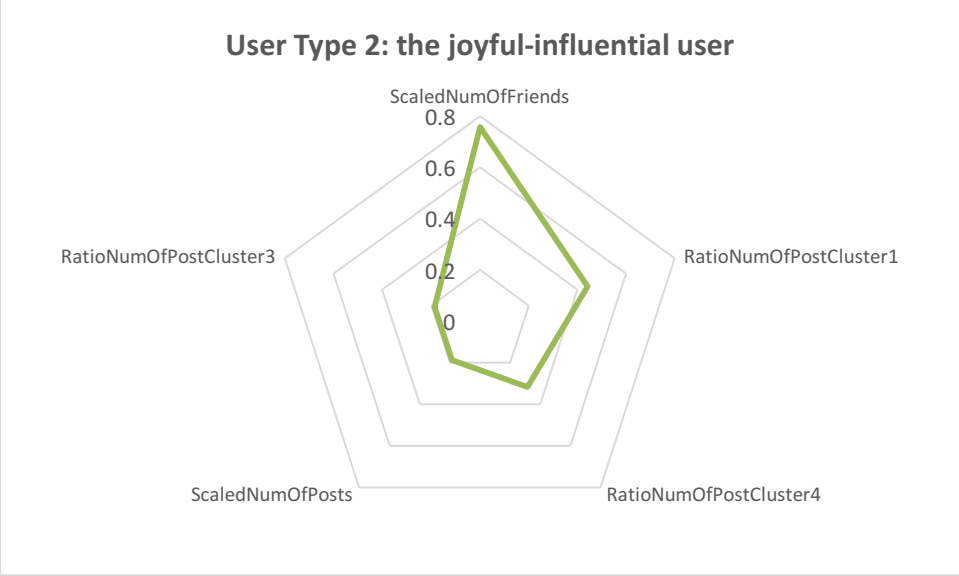


Figure 9 Top-5 most prominent characteristic for 'The joyful-influential users'

5.2.2.4 User Type 3: The cynical-lone users

The group of users who has less number of friends, profile photos, shared news, posts, self-uploaded photos/videos. As we can see from Figure 10, they passively either write posts on their wall or share posts from others. However, once they write posts, they often convey negative sense in short posts and maintain low interaction from others (high ratio value of PostCluster1). Here are examples of their post:

- Are all smart tvs worth their price, or are they just overpriced flatscreens with Netflix?
- Rain?...how unexpected..

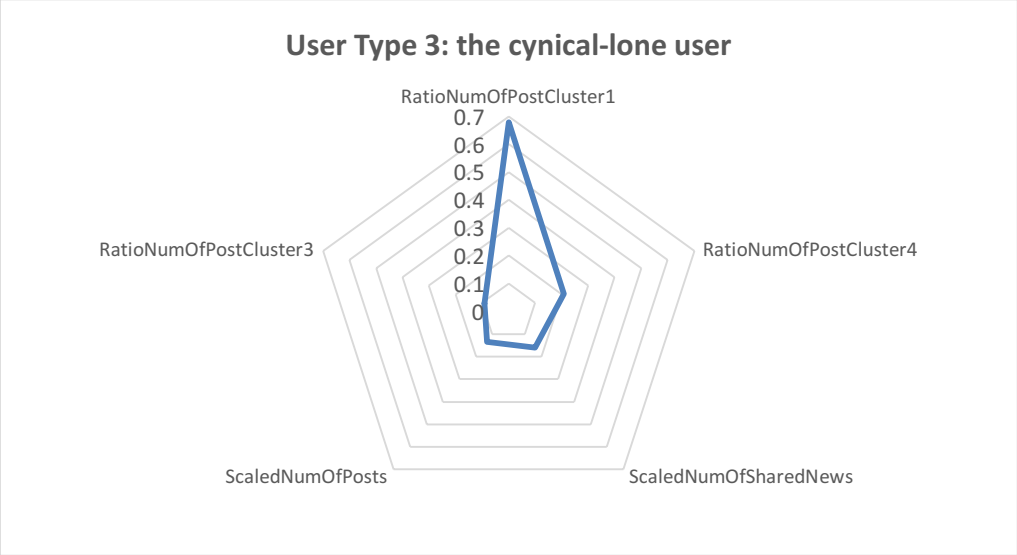


Figure 10 Top-5 most prominent characteristic for 'The cynical-lone users'

5.2.2.5 User Type 4: The joyful-lone users

As can be seen from Figure 11, this cluster represents the group of users who has high ratio value of PostsCluster4 with particular characteristics such as short-length posts, subjective and positive sense that are never shared by other, never receive likes, but sometimes receive comments. It seems like they often deliver the positive sense of posts but rarely gain feedback from others. Here are examples of their post:

- Had a great time with you guys. Awesome show, awesome free beers #therottengrapes #improv
- BEST DEAL AVAILABLE CALL AND RESERVE

Talking about their activity features, these users have the highest number of self-uploaded video. However, since this feature is not much differentiating along different user types, so we do not count this feature as a typical feature to characterize users.

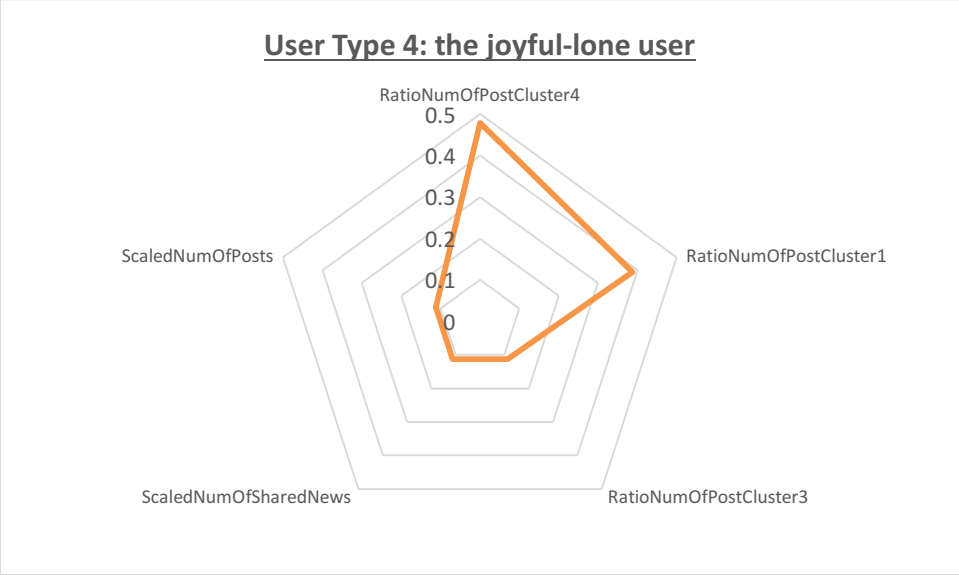
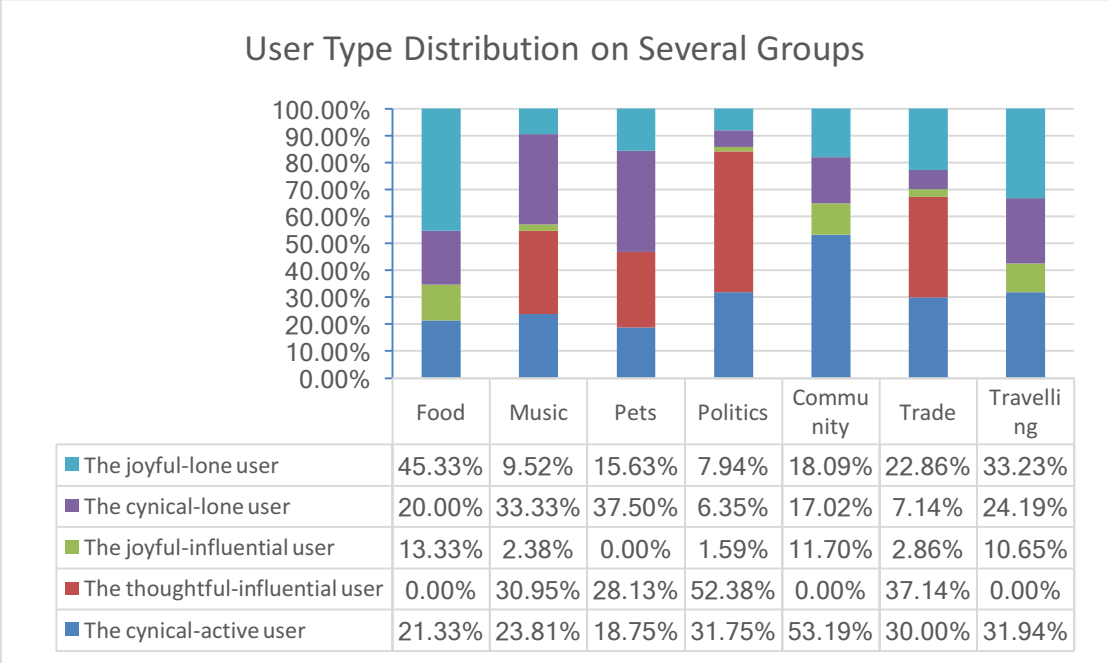


Figure 11 Top-5 most prominent characteristic for 'The joyful-lone users'

5.2.3 The Distribution of User Types in Interests Group

In this part, we explain how each Interest Group is characterized based on the distribution of user types. We collected Facebook Group from several topics including Community, Food, Music, Pets, Politics, Trade, and Travelling. Facebook Group is a place for Facebook users to interact, communicate, and join a discussion about the common interests.

Our aim is to evaluate the moods of Facebook Groups whether the Group is occupied with a number of particular user type and to conclude similarities of posting behaviors in several Facebook Groups. Noted that in this research, a single user is associated with only one Facebook Group.



	Food	Music	Pets	Politics	Community	Trade	Travelling
The joyful-lone user	45.33%	9.52%	15.63%	7.94%	18.09%	22.86%	33.23%
The cynical-lone user	20.00%	33.33%	37.50%	6.35%	17.02%	7.14%	24.19%
The joyful-influential user	13.33%	2.38%	0.00%	1.59%	11.70%	2.86%	10.65%
The thoughtful-influential user	0.00%	30.95%	28.13%	52.38%	0.00%	37.14%	0.00%
The cynical-active user	21.33%	23.81%	18.75%	31.75%	53.19%	30.00%	31.94%

Figure 12 The distribution of user types on several groups

As presented in Figure 12, almost half of the percentage of Food Group is occupied with ‘the joyful-lone users’. Music Group is more various in users’ distribution. 33.33% of the group belongs to ‘the cynical-lone users’, followed by ‘the thoughtful-influential users’. Pets Group is populated with ‘the cynical-lone users’ and ‘the cynical-active users’. Unlike Pets Group, Politics Group is mostly occupied with ‘the thoughtful-influential users’ and ‘the cynical-active users’. In Community Group, more than half of the percentage is occupied with ‘the cynical-active users’. Since this user type leads in number, we can say that the characteristic of the ‘the cynical-active users’ resembles in Community Group.

The Trade Group is populated with ‘the thoughtful-influential users’ and ‘the cynical-active users’. Both user types are characterized this group and Politics Group, but Politics Group possesses more users with these user types rather than Trade Group. Travelling Group is simply occupied with various user type. The percentage of ‘the joyful-lone users’ and ‘the cynical-active users’ are dominant in this group.

6 Conclusion and Future Work

We presented a work on clustering SNS users from their publicly available posts that attempts to discover user type based on sets of behaviors that commonly co-occur through their online record activities on SNS, especially in Facebook and to characterize its relationships with particular interest groups.

We discovered the five distinct user types that define online activities happened on Facebook and summarized the characteristics as follows: (1) 'the cynical-active users' represent the group of users who actively writes posts and shares news from others but seldom receives feedback; (2) 'the thoughtful-influential users' typify the group of quite popular and interactive users who often writes formal and serious posts that mostly earn feedback; (3) 'the joyful-influential users' belong to the group of users who often makes friends, writes positive posts that frequently gain interaction; (4) 'the cynical-lone users' represent the group of users who has the least number of activities on Facebook, once they write posts, they convey negative sense that maintain low interaction; and (5) 'the joyful-lone users' define the group of users who often delivers the positive sense of posts but rarely gain feedback from others.

We examine the distribution of user types on each interest group. Community Group is filled with people from 'the cynical-active users', while Food Group is mostly occupied with users from 'the joyful-lone users'. Music Group is filled with users from 'the cynical-lone users' and 'the thoughtful-influential users'. Pets Group is populated with users from 'the cynical-lone users' and 'the cynical-active users'. Politics and Trade Group shares the similar distribution of user type. Both are occupied with 'the thoughtful-influential users' and 'the cynical-active users', but Politics Group possesses more users with those type rather than Trade Group. Travelling Group is simply occupied with various user type. The percentage of the 'the joyful-lone users' and 'the cynical-active users' are dominant in this group.

There are still rooms to improve our proposed method. For the future work, we plan on clustering users with other clustering method and utilize more features.

7 Acknowledgement

Alhamdulillahhirabbil'alamin. All praise to Allah SWT who granted me the courage, capability, and opportunity to complete my master degree. It has been a long journey with hardship and many turns. I have received a lot of support, guidance, and kindness from people whom I met throughout this journey and for that, it is a pleasure for me to express my gratitude for them.

First of all, I want to show my sincere gratitude to my supervisor Professor Mizuho Iwaihara, for giving me the opportunity to do research in the Data Engineering laboratory. I highly appreciate his supervision, constant support, guidance, and encouragement throughout my research. I admire his hard work and your fast response in correcting my thesis research. Without your selfless assistance, strict supervision and impressive patience, I could not finish my master study and complete my master's thesis. Thank you very much.

I would like to thank *Lembaga Pengelola Dana Pendidikan*/Indonesian Endowment Fund for Education (LPDP) that provides me with the financial support till the completion of my master degree.

I also want to thank all members of Iwaihara for who gave me a lot of cheers, kindness, and support.

Finally, I want to deliver my gratitude to my parents for their unconditional love and continuous support. I would not have made it this far without your support.

References.

[1] Anon, Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/# [Accessed January 14, 2017a].

[2] Anon, Natural Language Toolkit — NLTK 3.0 documentation. Available at: http://www.nltk.org/ [Accessed January 14, 2017b].

[3] Anon, Selenium - Web Browser Automation. Available at: http://www.seleniumhq.org/ [Accessed January 14, 2017c].

[4] Anon, TextBlob: Simplified Text Processing — TextBlob 0.12.0.dev0 documentation. Available at: http://textblob.readthedocs.io/en/dev/ [Accessed January 14, 2017d].

[5] Bhattacharya, P. et al., 2014. Inferring user interests in the Twitter social network. *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14*, pp.357–360. Available at: http://dl.acm.org/citation.cfm?doid=2645710.2645765.

[6] Dunn, J.C., 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1), pp.95–104. Available at: http://www.tandfonline.com/loi/ucbs19%5Cnhttp://dx.doi.org/10.1080/01969727408546059%5Cnhttp://www.tandfonline.com/.

[7] Kim, J. et al., 2014. Extracting user interests on facebook. *International Journal of Distributed Sensor Networks*, 2014.

[8] Liang, Y., Sentiment Analysis on Publicly-Posted SNS Contents.

[9] Marquardt, J. et al., 2014. Age and gender identification in social media. *CEUR Workshop Proceedings*, 1180, pp.1129–1136.

[10] Mvungi, B. & Iwaihara, M., 2016. Estimating Purposes of Users in Social Networking Service Public Contents.

[11] O'Donovan, F.T. et al., 2013. Characterizing user behavior and information propagation on a social multimedia network. *Electronic Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2013*.

[12] Preoţiuc-Pietro, D. et al., 2016. Studying the Dark Triad of Personality through Twitter Behavior.

[13] Preoţiuc-Pietro, D. et al., 2015. Studying user income through language, behaviour and affect in social media. *PLoS ONE*, 10(9), pp.1–17.

[14] Volkova, S. et al., 2015. Inferring Latent User Properties from Texts Published in Social Media. *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*.

[15] Zhang, X. et al., 2007. A clustering algorithm based on mechanics. *Advances in Knowledge Discovery and Data Mining, Proceedings*, 4426, pp.367–378.n