Wolfgang Nejdl
Editor-in-Chief, The Journal of Web Science

Dear Dr. Wolfgang Nejdl:
Subject: Submission of revised paper #50

Thank you for your email dated 02-Aug-2017 enclosing the reviewers' comments. We have carefully reviewed the comments and have revised the manuscript accordingly.
In this letter, for simplicity, a comment ID such as A-1 is assigned to each of reviewers' comments (A-1 is 1st comment from reviewer A).
In our revised manuscript, we describe modification parts based on comments from reviewer A as colored in red, reviewer B as colored in blue, and reviewer E as colored in green. Each of modification parts has a corresponding comment ID at the top of the block. Our responses are given in a point-by-point manner. Also, deleted parts from our original manuscript are displayed in the strikethrough style.
After revising our original manuscript, we submitted it to English proofreading service to check English problems in our paper. According to the check, we corrected writing errors throughout the manuscript.

We hope the revised version is now suitable for publication and look forward to hearing from you in due course.

Sincerely,
Tomu Tominaga
Ph.D student at Osaka University

Responses to all reviewers

Thank you for taking a lot of time to review our paper. On behalf of the authors, I have answered each of your points below.

Before I show our response to your comments, there is one thing that I'd like you to know.

From all of you, we received five potential extensions in total: (1) estimation of user behaviors from types of profile images (from A), (2) text analysis (from B), (3) analysis on demographic information of users (from B), (4) evaluation on performance of multi-class classification (from B), and (5) extraction and analysis on image features with computational ways (from E). The associate editor of our manuscript, Dr. Ingmar Weber, asked us to include at least one in our revised manuscript among the five extensions. According to his indication, we addressed (1) and (4) because we concluded that these two issues are important and suitable for the scenario of our paper aiming to understand the relationship between user activities and profile images of Twitter users. Unfortunately, we could not examine (2), (3), and (5), but these extensions are also truly interesting. Therefore, we conduct all of these extensions in our future work.

Responses to reviewer A

Thank you for taking a lot of time to review our paper. On behalf of the authors, I have answered each of your points below.

Comment #1 (A-1)

> In Section 3.2. in Experiment 1, when coders classified 300 profile images, there was no mention about “out-of-service” users. But in Section 3.3 in Experiment 2, when coders classified 1200 profile images, 9.4% of them were out-of-service users. Why did Exp. 2 have out-of-service users but not Exp. 1? One possible reason would be that Exp. 2 was conducted later than Exp 1. If this is true, the paper needs to mention when the two experiments were conducted.

We found out-of-service users in Experiment 1. However, when we wrote the first version of this paper, we thought that the existence of “out-of-service” users was not so important because the Experiment 1 focuses on only live accounts and assesses the coincidence of people’s classification of profile images to the thirteen categories. However, we noticed that we need an explanation of “out-of-service user” from your comments in Experiment 1. Therefore, we add a simple explanation as shown below to Section 3.2 (A-1-3.2).

As pointed in (B-9), we noticed that we do not have to include “out-of-service” users in the examination in Section 3.3, because this section examines the ratio of profile images which can be categorized to our thirteen categories only for live accounts. We excluded the bar of “out-of-service users” (O.S.) from the graph (Figure 2) and modified the description about “out-of-service users” in Section 3.3 (A-1-3.3).

===

(A-1-3.2) From this large pool of unique users, we randomly picked 300 users after excluding “out-of-service” users, whose accounts of these users are frozen by Twitter (therefore they cannot use Twitter and we cannot get any information about them).

===

===

(A-1-3.3)

Original:

We found out-of-service users in our target users. User accounts of these users are frozen by Twitter, therefore they cannot use Twitter and we cannot get any information about them. They are classified into “O.S.” category shown in the right of this figure.

We mean “others” category by “Ot.”.

Of the 1200 users, 113 users are out of service in Twitter. Therefore the number of target profile images is 1087, which still meets the condition n>=1067.

Revised:

We found 113 out-of-service users in our sample and excluded them from this analysis. Finally, we obtained category labels for the 1,087 profile images, meeting the required sample size, n>=1067.

===

Comment #2 (A-2)

> Related to the above point (A-1), how often do Twitter users change their profile images? Also, when they replace their profile image, is the new profile image in the same category as the previous image—do users have consistent categorical patterns for profile image selection? Given that the work is using one-month Twitter data for their analyses, the results may not be affected by it, but I think it should be mentioned and discussed.

As the reviewer pointed out, there is a possibility that they changed their profile images from one category to another category. Also, we did not actually inspect whether or how often they changed their profile images. Instead, according to investigation by Whitty et al. (2017), there were only approximately 10% Twitter users who changed profile images once a month or more.

I add description about this issue as below to Section 3.2.

===

(A-2) During the period cited, users had an opportunity to change their profile images. However, Whitty et al. (2017) showed that only 10.8% of Twitter users changed their profile images once per month or more. Our data for this analysis were gathered for a period of one month; therefore, we assume that our results are not largely affected by the image-change phenomenon.

===

Comment #3 (A-3)

> In Section 3.3, “Therefore the number of target profile images is 1087, which also meets the condition n>= 1067.” What does this "condition" mean and how does it relate to the validation process?

In order to ensure statistical validity, we calculated the number of users to be coded using the website: https://www.surveysystem.com/sscalc.htm. Here, we let the confidence level be 95%, the confidence interval be 3.0, and population be 20,833,001, which is the number of tweets in our large pool, then we got 1067 as a sample size needed. This meant “the condition”. We should have mentioned what is the condition and how we calculated the number 1067. Thus, we add the description as shown below to Section 3.3. to briefly explain this.

===

(A-3) Before coding the profile images, we considered the statistical number of images needed, calculating a sample size of 1,067. The formula used is shown below.

((formula))

In this calculation, we set the confidence level to 95%, giving us Z=1.96; the confidence interval is 3.0, giving us c=0.03; and the population N is 20,833,001, which is the number of tweets in our large pool. Moreover, we set p=0.50, the most general setting.

===

Comment #4 (A-4)

> In Section 5.1. Data collection, "…, we randomly picked up 100 users for each category from the large pool …," how was this random selection made? For example, did authors or other coders manually examine the users/images from a list of users?

We should have explained this procedure in detail. In this process, we firstly picked up a target user randomly from the 4M users in our large pool. Next, a profile image of the target user was shown to coders, then the coders were asked to categorize its image into one of the 13 categories. We conducted this classification process until the number of users classified into each category reached 100.

We change description as shown below.

===

(A-4)

Original:

Firstly, apart from experiments in Subsection 3.2 and 3.3, we randomly picked up 100 users for each category from the large pool of the 4M users.

Revised:

First, we pick a target user randomly from the pool of four million users. Next, a target user's profile image is shown to the coders, who then classify it into one of 13 categories. The process is repeated until the number of users classified into each category reaches exactly 100.

===

Comment #5 (A-5)

> In Section 5.1. Data collection, authors mentioned "Table 2 summarizes medians and standard deviations …," but Table 2 does not include standard deviations.

We add standard deviations to Table 2.

Comment #6 (A-6)

> One suggestion for the future work: predicting user activities from one’s profile image selection. While it is interesting to predict users profile image category based on their activities, the opposite direction prediction can also be meaningful — it could have practical implication for solving a cold-start problem. For example, if one sets his/her profile image as Otaku category, then Twitter can use that information to suggest who to follow.

Thank you for your precious suggestion. I agree with the reviewer because the prediction of user activities from types of profile images would be also practical. According to the reviewer’s proposal, we add analysis (A-6-5.2.3) and results and discussion (A-6-6.3) about estimation of user behaviors from the types of profile image.

Also, we add brief conclusion about this to Section 8.

===

(A-6-5.2.3) Prediction of user activities from profile image categories

The purpose of this analysis is to obtain insights about the extent to which we predict user activities from users' profile images. To this end, we examine mean values and confidence intervals of user activities per categories of profile images. Results show the mean and an interval estimated by observed user activity data based on profile images. Therefore, we can estimate how users behave from their profile image selection. In this analysis, we use 95% confidence intervals.

===

===

(A-6-6.3) Prediction of user activities from profile image categories

In this section, we discuss the extent to which we can predict user activities from categories of target user profile images. To this end, we calculate means and confidential intervals (CI) of user activities per the 13 categories. Figure 5 shows the mean values (plots) and 95% CI (dotted lines) of target user activities in each category, and aligns the 13 categories from small (left) to large (right), in terms of CI range. The y-axis fits the range of mean values; confidence intervals are partly cut off.

Results demonstrate that the CI of user activities for categories tend to reflect larger ranges if categories show higher mean values in these activities. In other words, it is difficult to predict frequent user activities with high probability. For example, in Figure 5(a), “oneself” (On), “different person” (Dp), “letter” (Le), or “hidden face” (Hf) categories have higher mean values and wider CI ranges of FF. Compared to these four categories, “logo” (Lo) has a smaller CI range and a higher mean value. Thus, predicting FF of “logo” users ought to be easier. For “associate” (As), “character” (Ch), or “self portrait” (Sp) categories, mean values of FF are lower, which correspond to lower CI ranges. Twitter users in these three categories tend to balance their follower-to-followee ratio. Otherwise, these users are likely to engage in reciprocal relationships, following someone after the person follows them, or vice versa. Possibly, we predict FF of users in these categories with relatively high probability.

Additionally, as shown in Figure 5(b), 5(d), 5(f), and 5(g), Rtw, Rrted, Rurl, and Rhash also demonstrate that larger mean values are likely to reflect wider CI ranges. For these activities, "associate" (As) category shows the smallest CI ranges. As with FF, we might predict user activities of "associate" users with high probability. "Otaku" (Ot) shows the highest mean value and the widest CI range of Rtw, the daily frequency of posting tweets (42.70+/-7.63, Figure 5(b)). Compared to other categories, highly accurate prediction for posting frequency of "otaku" users is relatively difficult. However, on average, we estimate that "otaku" users post tweets once or twice per hour. Concerning Rrted, Rurl, and Rhash, estimating the frequency of user activities for "logo" (Lo) and "letter" (Le) users ought to be easier than FF, because their CI ranges are smaller.

We also find that there are no large variances in CI ranges of Rrt and Rrep, as shown in Figure 5(c) and 5(e). Interestingly, "associate" (As) category shows the highest mean value and the smallest CI range of Rrep (0.63+/-0.04). This result shows that most "associate" Twitter users massively engage in replying activities. Therefore, if we detect that a new user chooses a group photo as a profile image, the user might often reply to other users. Specifically, the estimated frequency of replying is approximately five-to-seven times per day, which is calculated from statistics shown in Figure 5(b) and 5(e).

Our results can be used to tackle cold-start problem (Schein et al., 2002) in Twitter user recommendations, because when a new user sets a profile image, we can estimate user activity based on the given profile image. For example, if a new user is found to select an "associate" photo as a profile image, the user would be comfortable if users who prefer communication with the associate user such as real-world friends or family members were recommended. New "logo" or "letter" users will frequently post tweets with URL links or hashtags, and will reply two-to-four times per day. Therefore, it might be useful for new "logo" or "letter" users to receive recommendations to connect with similar users; they would learn faster how to promote their content from the behaviors of the recommended users. Moreover, they do not prefer in-person communication on Twitter. Therefore, their reply messages to someone should not be promoted so that these messages do not automatically appear in timelines of non-follower users.

===

Responses to reviewer B

Thank you for taking a lot of time to review our paper. On behalf of the authors, I have answered each of your points below.

Comment #1 (B-1)

> The analysis of the paper is overall well executed. However, there are many grammar errors and strange formulations, so the paper must be thoroughly proof read by a native speaker of English or alternatively, provide line numbers with the submission so that the most obvious mistakes can be pointed out.

As the reviewer pointed out, our manuscript should have proofread by a native English speaker. In order to fix our grammatical errors, we submit our revised paper to English proofread service.

Also, I provide line numbers with the revised manuscript.

Comment #2 (B-2)

> The future research directions identified are all very natural and interesting problems. I would have preferred if any of them (e.g. text analysis as this seems the most straightforward) would have been included in this version of the paper. Another natural direction for this work would be to create models that can automatically predict the 13 categories of images in unseen profile images. This would allow the cross-cultural analysis to be done more easily. Another outcome of this paper, which is only implied but not mentioned explicitly is that selecting to annotate only users who present themselves in the profile image leads to a biased sample of Twitter activities. This is an important point which should be taken into consideration by other lines of research.

Thank you for your precious proposals. I agree with the reviewer in that text analysis and prediction models would give us more insights about difference in user behaviors according to types of profile images. In this study, we focus on simple user activities of Twitter users and show how effective these activities are to predict types of profile images as a first step of this project. The advantage of using simple activity features but contents features is high reproducibility. We did our best in examining the relationship between profile images and user activities using these simple features. For this position, we conducted additional analysis on multi-class classification issue as described in (B-4).

Concerning to the issue on a biased sample, unfortunately, we cannot completely interpret the intention of the reviewer. In this study, not only users who present themselves such as oneself or associate users but also users who set default images as profile images are included in our target users; therefore, we believe that our dataset is not largely biased in terms of this point.

Comment #3 (B-3)

> One main concern I have with the results of this paper is that they may (and some look like) being driven by demographic preferences; for example, users with a profile picture of themselves are probably more likely to be older, which may correlate to posting more URLs. Even if this is not performed in the paper (albeit done in Liu et al 2016), this limitation and potential confound should be explicitly stated in this paper.

I agree with the reviewer. We need to discuss effects of demographic information such as age or gender. I add the description as shown below to Section 7.

===

(B-3) Moreover, our results may be influenced by target user demographics. For example, users who show themselves in their profile images are more likely to be older, inferring a high likelihood of posting URLs. There may be more females in the “associate” category, which might correlate to frequent replies. In our future work, we will inspect the effects of users' demographic information.

===

Comment #4 (B-4)

> Another results I would like to see in the paper is the performance of a 13-way classifier that identifies the class of a user from the activity features.

Thank you for additional proposal. This potential outcome is suitable to the scenario of this paper; therefore we conduct this classification and show the results in the revised manuscript. Concerning to the 13-way classification, I add a new section “(B) 13-way classifier” under Section 5.2.2 to explain method of this classification (B-4-5.2.2.B). Also, I add a new subsection 6.2.2 to show the results of this classification (B-4-6.2.2). According to this changes, we add section titles “(A) 2-way classifier” under Section 5.2.2 and “6.2.1 2-way classifier” under Section 6.2.

In addition, we add brief conclusion about this to the Section 8.

===

(B-4-5.2.2.B)

(B) 13-way classifier

In addition to 2-way classifiers, we aim to build 13-way classifiers that predict the types of users' profile images based on user activities. For building prediction models, we use the same machine learning techniques as before.

Here, we adopt a 5-cross validation so that 20 users from each category are selected as a test set, because 13 classes are prepared and the number of users selected from each category should be larger than the number of classes. In addition to the 2-way classifier, we train classifiers using training sets as tuning parameters to the best performance. After training classifiers, we evaluate classifier performance. We show F-measures for all classes and a macro-averaging F-measure, which is a mean of F-measures for all classes, to examine overall classifier performance.

===

===

(B-4-6.2.2)

6.2.2 13-way classifiers

Figure 4 shows F-measures of each category by 13-way classifiers according to the models. In Figure 4, a cumulative bar located at the right, named “Ave.”, shows a mean of F-measures of each model. This is called macro-Fscore (Sokolova and Lapalme, 2009), which represents overall performance of multi-class classifiers. The baseline is a random classifier and its F-measure is 1/13 (0.077), because the number of classes is 13 for each classifier and we prepared the same number of users in each category (both of precision and recall of random classifiers are 1/13). Similar to the 2-way classifiers, we show the F-measures calculated by the 13-way classifiers as mul-f^m_c, where m and c are mean types of machine learning models and categories of profile images, respectively.

“Associate” (As), “otaku” (Ot), “default” (De), and “logo” (Lo) categories show better performance. Among all models of all categories, the random forest model for “associate” shows the best performance (mul-f^rf_As=0.422). As mentioned before, these categories reflect unique user activities and combinations. Thus, 13-way classifiers also easily identify users per their categories.

Concerning to overall performance of all models, random forest is the best (mul-f^rf_c =0.209), as well as for the 2-way classifiers. As mentioned in 6.2.1, this machine learning technique directly improves classification results based on the Gini index. Thus, it can consistently show better performance over the other machine learning techniques, plus, it tackles multi-classification problems.

Compared to the 2-way classifiers, differences in performance among categories are relatively larger. If a category does not possess many unique user activities, classifiers cannot obtain clues to know whether users belong to a given category. Thus, users in a category that does not perform many unique activities are likely to be predicted as users from other categories. For example, the number of users predicted as belonging to target categories is sometimes 0 for “hidden face,” “scene,” or “different person.”

In contrast to the two-way classifiers, “otaku” shows better performance than “associate” using logistic regression and SVM models (mul-f^lr_Ot, mul-f^lr_As=0.380, 0.365, mul-f^svm_Ot, mul-f^svm_As=0.393, 0.384). Examining the prediction results in detail, we find that the number of predicted users in “associate” tend to be larger in these two machine learning models than in the random forest models. This causes model precision to be low. Therefore, we find lower F-measures in “associate.” At this stage, we cannot provide insights into the reason why precision is worse. However, we will tackle this issue in future work.

===

Comment #5 (B-5)

Section 2, paragraph 2: 'two types of personality traits' -> personality traits are not of different types, but calculated with different methods

I fixed the sentence as shown below.

===

(B-5) Cristani et al. (2013) and Segalin et al. (2016) showed that features of user images labeled as “favourite” correlates with personality traits, which are calculated in two ways: self-reported personality and rater-reported personality.

===

Comment #6 (B-6)

Figure 1: please be sure that you have the rights from the users to publish these photos as part of the paper

I changed profile images in Fig. 1 of users who permit us to publish these photos in this paper.

Comment #7 (B-7)

Section 3: I would rather not have the two items described as research questions - these are just validation steps for the user categorization and the label acquisition validity

As the reviewer pointed out, these are not research questions but just validation steps. According to this comment, I change the description as shown below.

===

(B-7) The categories in Table 1 are empirically established, therefore we go through two validation steps as follows. First, we validate the rate of coincidence of people's classifications of profile images. Then, we validate the 13 categories’ coverage ability for general profile images on Twitter. These steps are called coincidence and coverage steps.

===

Comment #8 (B-8)

Section 3.3: who are the coders that performed the labeling

The coders are students in graduate school at Osaka University. I add description to explain who the coders are to Section 3.2 and 3.3.

===

(B-8) First, we invited four coders: graduate students at Osaka University, receiving compensation for this experiment.

===

Comment #9 (B-9)

> Section 3.3: the 'O.S.' users should be removed beforehand as they are irrelevant to the analysis and can be identified in advance

As described in (A-1-3.3), we shortly explain the existence of out-of-service users, and delete the bar of these users in Figure 2.

Comment #10 (B-10)

> Section 3.3: the description for how .914 was computed is not needed

I delete this description.

Comment #11 (B-11)

> 'Egg' has recently been replaced by an abstract shape of a person as the default for Twitter; may be better to rename the class

I changed the name of this category from “egg” to “default” throughout this paper.

Comment #12 (B-12)

> Section 4: This section is very verbose for describing a few simple features. I would like this to be described more concisely and perhaps summarized through a table that allows a quick look-up of different variables

For allowing readers to look user activities in brief, I add subtitles at the beginning of each section about the user activities.

Comment #13 (B-13)

> Section 4, paragraph 8: the assumption for what is performed with an URL ('detailed information') is quite strong and not often accurate. Many times URLs link to a video or photo

As the reviewer pointed out, links to a video or a photo are also included. Thus, I change the term “more detailed” to “additional” because we believe that links, including videos and photos, provide more information to plain texts in tweets.

Comment #14 (B-14)

Section 5.1: How are the 100 images/category identified if the class distribution in the real-world in different? Wouldn't the coders annotate more images until 100/category is reached, resulting in some categories having more users.

We should have explained this process more specifically. As describe in (A-4), in this procedure, we picked up a target user randomly from the 4M users in our large pool. Next, a profile image of the target user was shown to the coders, then we asked the coders to classify profile images into the 13 categories. Here, coders repeated this classification until 100 users are classified into each category. If the number of users in a specific category reaches 100, we asked coders to stop classification of profile images into the category (its profile image is discarded).

I add the explanation as shown below following the part of (A-4).

===

(B-14) In case that the number of users classified into a specific category reaches 100, we asked the coders to discontinue classifying profile images into the category and discard the images.

===

Comment #15 (B-15)

Section 5.1: Twitter REST API -> Twitter Search API

According to Twitter Inc. (specifically in this page: https://dev.twitter.com/rest/public), Twitter Search API is one of Twitter REST APIs. We used REST APIs to get data related to followers, followees, or tweets of our target users.

Comment #16 (B-16)

Table 2 does not have standard deviations as mentioned in the text from Section 5.1

We add standard deviations to Table 2. As the reviewer A also pointed this, the text color is red.

Comment #17 (B-17)

Section 5.2.1: one vs. all classifiers are a generally well-known name, no need to go into details

I delete the explanation “For instance, let the target category be “oneself”, users in the category are labelled as positive and other users are labelled as negative”.

Comment #18 (B-18)

> Table 2 has some strange numbers: for example 'Egg' has 0.0 for 5/7 features. Also, the Rhash has very low values across the board, which is not the case usually for tweets.

There are a lot of novice users on Twitter in egg (default) category. In our dataset, most of egg (default) users have not posted tweets yet as shown by a median of Rtw of these users. Also, it is median values that Table 2 shows; therefore, considering that Rhash is tend to be low in this table, hashtags may be frequently used by only a part of Twitter users.

Comment #19 (B-19)

> Section 6.1, paragraph 4: the length of the tweets hypothesis can be easily checked and the features computed, rather than speculated upon.

I add a new table showing medians of average length of users’ tweets according to the 13 categories, and change the description about word counts as shown below.

===

(B-19)

Original:

Manually checking, we found that their tweets tend to have fewer words than tweets from users in other categories (among otaku users, median of word count per tweet is 25, which is the minimum number of all categories).

Revised:

Here, we study the average tweet word count of users in the 13 categories. We exclude any usernames, hashtags, and URL links from users' tweets. We summarize the results in Table 4, showing medians of the average word count per the 13 categories. “Otaku” has the smallest median: 24.70.

===

Comment #20 (B-20)

> Figure 3: will be better to have as a single feature with stacked bars for each class for easy comparison across the three methods

As the reviewer pointed out, we should have designed this figure to let readers to compare performance of each model easily. In the revised manuscript, we show the results of cumulative F-measures with stacked bar plots for easy comparison (Figure 3 and Figure 4).

Comment #21 (B-21)

Section 6.2, paragraph 4: the explanation at the end is probably not valid as 'otaku' and 'egg' represent only 2/12 cases in the 'not' category

By the sentence “Accordingly, it was difficult for the models to distinguish hidden face users from otaku users, and animal users from egg users”, we would like to mean that one of the reasons behind worse performance of hidden face and animal users might be that otaku and egg (default) users are also significant in Rtw and Rrted; therefore the machine learning models might be confused to distinguish hidden face or animal users from otaku or egg (default) users.

As the reviewer intended, the description in the original manuscript was too strong; therefore I fixed this description as shown below.

===

(B-21) Moreover, “hidden face” and “animal” users perform less unique activities than “otaku” and “default” users. Therefore, “hidden face” and “animal” categories might pose difficulties for the machine learning models when distinguishing “hidden face” users from “otaku” users and “animal” users from “default” users.

===

Comment #22 (B-22)

Some of the English mistakes from the paper that should be addressed:

(1) Section 2, paragraph 3: 'less likely to be active' -> mention the personality trait rather than 'active'

(2) Section 2, paragraph 4: 'look fat' -> 'are overweight'

(3) Section 3.1: 'images are substantially corresponded' ?

(4) Section 4, paragraph 1: strange construction of the first sentence

(5) Section 4, paragraph 1: 'usage tendency' -> strange usage of tendency; exists in more places

(6) Section 4, paragraph 4: 'expanding information' -> strange and innacuratte

(7) Section 4, paragraph 7: 'a lot of users' -> 'more users'

(8) Section 5.2 title: 'Analysis method' -> just 'Analysis'

(9) Section 5.2, paragraph 1: 'they are belong'

(10) Section 5.2.2, paragraph 2: 'and their user activities' -> out of place here, delete

(11) Section 5.2.2, paragraph 3: 'for the training dataset, we train the classification models' -> ?

(12) Section 6.1, paragraph 3: 'post tweet' -> 'post more tweets'

(13) Section 6.2, paragraph 4: 'this leads the worse performance'

Thank you for your detailed comments on our English mistakes. To fix English mistakes throughout the paper, we use English-proofreading service. After the paper is edited by the service, I fixed again our manuscript according to the reviewer’s comments if the mistakes still remain.

Responses to reviewer E

Thank you for taking a lot of time to review our paper. On behalf of the authors, I have answered each of your points below.

Comment #1 (E-1)

> However, the choice of annotating by hand the 13 categories of objects observed in the users' profile images makes the technical contribution of the paper quite weak. As mentioned also by the authors there are several works such as the ones by Segalin et al., Cristani et al., Wei and Stillwell, Celli et al. (not mentioned by the authors) that use standard computer vision approaches to extract image characteristics from profile pictures. My suggestion to the authors is to improve the paper adding this technical contribution. In this way, they can also overcome the time-consuming activity of asking coders to annotate images and thus they can obtain a larger dataset for their analyses.

As the reviewer suggested, types of profile images in our dataset are annotated by humans; therefore, it takes time to obtain a large dataset. I agree with the reviewer in that using computer vision techniques to extract image features is also valid. In this study, we aim to find categories of Twitter profile images in Japan, and to examine user activities according to the categories. To the best of our knowledge, no study examines what kind of categories of profile images in Japan there are. Thus, as a first step, we conducted manual annotation of profile images.

To add technical contributions, we conducted further analysis on multi-class classification as described in (B-4) and on prediction of user activities from profile image categories as described in (A-6).

Comment #2 (E-2)

> Additionally, the paper requires an intensive proof-reading from a native speaker. It's full of typos and sentences not well formed. To give some examples:
> - title: through - not though
> - Section 1: Twitter are relatively smaller than those ... not small than
> - detailed information - not detail information
>
> Again, formula 1 seems wrong or at least it's not the ratio of followees to followers but the ratio of followers to followees ...

Thank you for pointing our English mistakes. To fix English mistakes throughout the paper, we use English-proofreading service. After the paper is edited by the service, I fixed again our manuscript according to the reviewer’s comments if the mistakes still remain.