“Teens are from Mars, Adults are from Venus”: Analyzing and Predicting Age Groups with Behavioral Characteristics in Instagram

Kyungsik Han[1], Sanghack Lee[2], Jin Yea Jang[2], Yong Jung[2], and Dongwon Lee[2]

[1]Pacific Northwest National Laboratory, USA, [2]The Pennsylvania State University, USA

kyungsik.han@pnnl.gov, {sxl439, jzj157, yuj114, dongwon}@psu.edu

ABSTRACT

We present behavioral characteristics of teens and adults in Instagram and prediction of them from their behaviors. Based on two independently created datasets from user profiles and tags, we identify teens and adults, and carry out comparative analyses on their online behaviors. Our study reveals: (1) significant behavioral differences between two age groups; (2) the empirical evidence of classifying teens and adults with up to 82% accuracy, using traditional predictive models, while two baseline methods achieve 68% at best; and (3) the robustness of our models by achieving 76%—81% when tested against an independent dataset obtained without using user profiles or tags. Our datasets are available at: https://goo.gl/LqTYNv

CCS Concepts

• Human-centered computing~Social media
• Human-centered computing~Social network analysis
• Human-centered computing~Social networking sites
• Human-centered computing~Empirical studies in collaborative and social computing

Keywords

Teens in social media; behavioral patterns and detection of teens in social media; comparative analysis

1. INTRODUCTION

We live in a digital era where technologies have enabled us to create and interact with a variety of information based on our needs and situations and to develop and maintain social relationships with other people, including family members, friends, co-workers, and even strangers. Social media has greatly re-shaped the way people are present online by providing a channel and space to socially connect with others, share one’s life updates, express one’s opinions, and receive up-to-date information on topics that they care about and are interested in [20]. It allows people to share content in text, image, or video as well as engage with others by offering simple and interactive features, such as Likes or comments.

There exist many demographical and psychological factors that can be considered for the study of social media use and engagement. Especially when it comes to age, the influence of social media on teens is significant [3,5,15], because they not only grow up with an abundance of communication technology but also are the heaviest users among all age groups [33]. The 2015 Pew report [22] indicates that 92% of teens go online daily, where 24% of them answered “almost constantly,” and 71% of teens use more than one social media platform, showing high platform diversity.

This heavy and diverse use and engagement in social media provides an opportunity to articulate teens’ online behaviors from their real usage data. This is important, because social media has already been indispensable part of their life, and its usage directly shows many aspects of themselves such as their personal interests, goals, situations, moods, concerns, etc. Given that teens are in the developmental stage where they experience rapid physical, emotional, intellectual, and social changes, it is our role, as researchers, to understand how they behave and engage in online space and to think about how social media can be used as an effective channel for teens’ development.

With such motivations, as the first step, we investigate the behavioral patterns of teens in social media. We especially study Instagram with the following research questions: (RQ1) *what are the behavioral characteristics of different age groups in Instagram?* (RQ2) *can one predict age groups based on the studied behavioral characteristics?* To answer these questions, we employ a comparative analysis method from teen and adult groups. This paper leverages our preliminary findings [16] and makes the following contributions:

- We address a growing concern in social media research on the use of a “biased” dataset and attempt to mitigate the issue by collecting two independently created and verified datasets, named Profile-based data (P-data) and Tag-based data (T-data), having a total of 20,000 teens and adults, and crosschecking the findings.
- Using both P-data and T-data, then, we analyze the behavioral characteristics of teens and adults through the lenses of three perspectives: *content-, interaction-, and relation-based activities*, covering how teens and adults create and use content, and interact with others in Instagram. We report that there are statistically significant differences in behavioral characteristics between teens and adults, especially in *content-* and *interaction-based* activities.

WebSci '16, May 22-25, 2016, Hannover, Germany
© 2016 ACM. ISBN 978-1-4503-4208-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2908131.2908160

- We demonstrate a possibility of building fairly accurate supervised learning models that can identify teens and adults with 82% accuracy (0.89 AUC) at best, while two baseline methods achieve 68% accuracy at best.
- We provide empirical evidence that activity-based variables used in the models are representative enough to differentiate two age groups by showing that our prediction models still yield up to 81% accuracy (0.85 AUC) when tested against a 3rd independent test dataset obtained without using user profiles or tags.

2. RELATED WORK

2.1 Understanding Teens and Adults in Social Media

Among other age groups, teens are believed to be the most engaged and adventurous in social media. Being acclaimed as digital natives [34], teens grow up with an abundance of communication technology and are believed to be more interested in using technologies than adults.

Teens appear to be early adopters and active users of social media [5]. For them, social media has become a new way to represent themselves [32], share their everyday activities and thoughts [8], establish and maintain social connections [27], and learn something new and useful [15]. According to prior research reports, teens tend to consider social media as an exciting opportunity for social interaction [15] and self-display [23], while adults may be more concerned about their information privacy in online disclosure [24]. Given the fact that socialization is an influential process in childhood and adolescence, interaction with their peers through social media plays an important role in teens' life and has an impact on teens' self-esteem and psychological well-being [38]. Their social needs drive their social media behaviors, and they actively interact in online to build and maintain connections with their peers.

Research has also presented social media use through a comparative analysis from different age groups. Pfeil et al. [29] showed that teens tend to express themselves and share their stories in online space more than older users (ages over 60). Teens also tend to use many social networking sites and maintain multiple forms of communications, compared to older users, because each channel offers different ways of interacting with others. Quinn et al. [35] found that younger users (ages 15-30, including teens) tend to use different social media features (e.g., updating their status, etc.) more than older users (ages over 50).

However, there have been much fewer quantitative studies on understanding teens and their behaviors in social media compared to those with qualitative approaches. Even those quantitative studies mostly rely on survey responses from small sample size (e.g., 50-300). In this paper, we present a possible way to overcome that issue.

2.2 Sampling and Generalization Issues

Much research has explored the impacts of social media on people and society; however, there has been a growing concern over using a limited method (e.g., single) for data collection and analysis. Many researchers have realized and claimed that social media research is not coherent with an established set of data collection, methods, evaluations, and documentation standards [6,40].

Recognizably, data collection is one of the most important processes for social media studies. Recently, much research has raised a potential issue of social experiments on limited datasets and warned a danger of having biased results from those experiments [30,31]. Apparently, many studies have relied on using public APIs, if available, or crawling web pages. However, the issue is about using a single method to collect datasets, which implies sampling and generalization issues on the study analyses and reports. Although it is challenging to completely resolve these concerns in social media studies, we attempted to mitigate them by using two data collection methods. Through this, we collected a total of 20,000 samples of teens and adults and investigated their behavioral differences in using and engaging in Instagram.

2.3 Modeling User Characteristics and Behaviors

We have seen an exceptional growth in social media data both in terms of their volume and diversity, as they include various types (e.g., text, photo, video), interactive activities (e.g., Likes), and social relationships (e.g., following others). As such, it is acknowledged that there is enough data to create reasonably accurate predictive models about users or social phenomena.

We are especially interested in the modeling and prediction of user characteristics and behaviors. There are indeed many studies in this topic. For example, Kosinski et al. [21] studied the application of Like activities to predict a number of user's personal traits such as gender, ethnicity, religion, political views, sexual orientation, intelligence, etc. from 58,000 volunteers on Facebook. They predicted individual psycho-demographic profiles from Likes through a regression model. They found that their model predicted homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases. Gilbert and Karahalios [11] presented a model built from a dataset of over 2,000 Facebook friendships. Each friendship was assessed by the metric called tie strength, which is based on 74 indicators. They found that a level of intimacy (e.g., days since last communication, number of friends, and intimacy words) is the most important factor to predict tie strength, followed by levels of intensity, duration, and social distance.

Similarly, we focus on understanding users (i.e., teens and adults) based on their social media use. However, our work is unique, because we present various types of activities and behaviors that clearly characterize teens in social media.

It is important to note that our objectives of this paper are not to simply compare teens with other specific age groups, but to explore the variables that reflect teens' social media activities and verify a possibility of building comprehensive models from those activity variables.

3. DATA COLLECTION

3.1 Instagram

Instagram is one of the most popular social networking sites [22]. Instagram users also show high engagement, as 59% of them are on the platform daily [7]. Because of its high popularity, there have been a growing number of studies on Instagram; for example, exploring the relationship between photo content and engagement [2], analyzing photo content and user types [14], exploring Like activities and networks [17], studying tag-based Like networks formed by users who share common tags [12], etc.

3.2 Challenges

To carry out a comparative study between teens and adults, ideally, one needs: (1) unbiased samples of Instagram users, and (2) ground truth about the ages of users in the samples. As neither is readily available to researchers outside of Instagram, we resort to the following methods:

- Collect 2 million Instagram users via followships from 1,000 random seed users.
- Using two different heuristics, create two separate datasets of 45,000 users each from 2 million users whose ages, either teen or adult, are accurately inferred and manually verified.

Using the Instagram API, we collected data for six months (October 2014-March 2015). With the data, we defined our target user populations by following Erikson's eight stages of psychosocial development [9]. We chose two age groups, "adolescence" (13-19 years old) and "early adulthood" (20-39 years old), because they are primary user populations in Instagram [7]. However, we had to add a 10-year gap between two age groups, due to a limitation of current age-detection tools, which will be explained in the next section, to warrant that all samples correctly represent each group. As a result, the followings are two user groups studied in this paper. For simplicity, we refer to adolescence as *teens* and early adulthood as *adults*.

- *Teens*: people who are between 13 and 19 years old.
- *Adults*: people who are between 30 and 39 years old.

We retained only active users for the study by excluding users who did not post any photo between October 2014 and March 2015.

3.3 Two Heuristics to Determine Ages

Given an Instagram user, how can we accurately tell if the user is a teen or adult? Obtaining ages of users is not easy since many users set their age private. In addition, many social networking sites, including Instagram, do not ask users to indicate their age during registration. Therefore, we employed the following two heuristics (Figure 1) to obtain two age-inferred datasets, called P-data and T-data.

- **Profile-based age detection**. We previously reported a hybrid method to infer a user's age using profile photos and textual bio description [16]. Our method used a computer vision-based age detection tool, Face++[1], on profile photos and textual cues such as "I am a junior in high school" to identify age information. With this method, we were able to collect a large number of teens in Instagram. However, since Face++ estimates age in a range (e.g., 18±5), many users lie in multiple age groups, e.g., adolescence (13-19) and early adulthood (20-39). Further, even during a manual verification process (will be described later), it was difficult for human judges to distinguish users in their late 10s and early 20s by looking at their profile photos. Therefore, we intentionally placed a 10-year gap (the maximum range returned by Face++) between two age groups to reduce ambiguity. Finally, we collected 30,000 teens and 30,000 adults (unverified samples).
- **Tag-based age detection**. As a new data collection method, we exploited a set of specific tags in photos, which imply age information. For example, #sweet16 or #my18birthday would imply that the user is a teen, while #30birthday or #hot35 would imply that the user is an adult. Finally, we identified 15,000 teens and 15,000 adults where none of them is included in the profile-based samples (unverified samples).

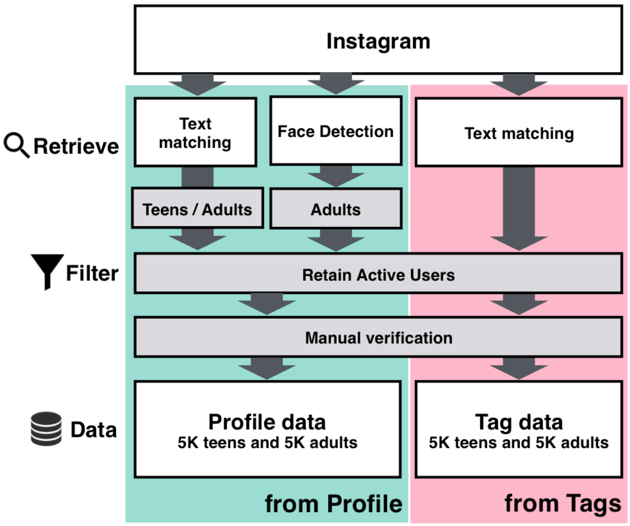

Figure 1. Data collection method. We collected a total of 20K teens and adults—5K teens and 5K adults from profile (i.e., P-data), and the same from tag (i.e., T-data) for data analysis.

After obtaining teens and adults from the heuristic methods, we further manually verified its correctness as follows. For each Instagram user, three human judges (i.e., authors) checked user's profile photo, bio description, and posted photos, and then labeled as a teen or an adult. We only kept users whose age group is unanimously determined. This manual verification process took about 3 days to complete all data in both profile-based and tag-based samples.

After human verification, we created: (1) **P-data**, 5,000 teens and 5,000 adults randomly chosen from the verified profile-based samples; and (2) **T-data**, 5,000 teens and 5,000 adults randomly chosen from the verified tag-based samples. Note that both P-data and T-data are mutually exclusive (i.e., no common users). It was our intention to have the same sample size in two datasets to minimize the bias caused from unbalanced data sets.

After we finalized a total of 10,000 sample users in each of P-data and T-data, we further collected various activities of those users including: (1) the number of photos, Likes, tags, comments, followers, and followings, (2) text information of tags and comments, and (3) locations of photos. Our final datasets are available at: https://goo.gl/LqTYNv

It is worthwhile to note that we paid special attentions to protect users' privacy from data collection to analysis. No user-sensitive information (e.g., name, gender, etc.) was retained, except Instagram user IDs that were needed in the Instagram API to collect user activities.

3.4 Benefits of Two Sample Datasets

Given two independently sampled datasets, we can analyze them either separately or together. When a similar result or trend is

[1] http://www.faceplusplus.com/

observed consistently in both datasets, one can be more confident about the finding. As we have P-data and T-data, here we describe our rationale of using them.

We first investigated whether distributions of each variable for each label in two datasets are different. We adopted a Kolmogorov-Smirnov (KS) two-sample test [25] and found that most variables (except teen's number of photos), used in our analyses, are significantly different ($p<0.001$) partly due to the size of data. Even when we substantially mitigated the effect of large data size on p-values by testing only on 100 random samples out of 10,000, we could still confirm that two data are quite different ($p<0.05$). Therefore, the distributions of each variable are statistically different, while having a similar shape in general.

We also compared two datasets with respect to the amount of difference, embedded in the variables of each dataset, between teens and adults. Those variables represent user's Instagram activities (e.g., number of photos, Likes, etc.), which will be described in the later section. We constructed predictive models to classify teens and adults and compared the performance of the models. Significant differences in their performances mean that two datasets are statistically different.

Table 1 illustrates different predictive ability of the models. There are notable differences when the model trained on T-data was tested on the same T-data or the other P-data. A careful examination reveals that the learned difference of two age groups in P-data also exists in T-data, since models trained with P-data yield similar performance on both P-data and T-data as test data. On the contrary, learned difference represented in T-data is rather unique, since it is not well generalized to P-data (i.e., performance gap between tests with P-data and with T-data given T-data trained models).

Table 1. Performance of Logistic Regression (LR) and Support Vector Machine (SVM) where Acc. and AUC stands for accuracy and Area Under Curve, respectively.

Training	Test	LR		SVM	
		Acc.	AUC	Acc.	AUC
P-data	P-data	0.720	0.783	0.752	0.826
	T-data	0.721	0.775	0.744	0.814
T-data	P-data	0.588	0.604	0.625	0.669
	T-data	0.803	0.875	0.824	0.892

These results justify the use of separate analysis from each dataset for hypothesis tests, which will reduce biases caused by using a single (merged) dataset. On the other hand, it was also found to be plausible to run analysis on a single merged dataset when researchers want to report descriptive statistics. Based on these insights, we employed both ways of handling data, depending on an analysis.

4. RESULTS

Our data analysis is based on three types of activities (i.e., content-based, interaction-based, and relation-based). Each type of activities represents different characteristics of social media use and engagement that Instagram users exhibit (Figure 2).

We define each perspective as follows:

- *Content-based activities* represent what posters added to their photos, such as tags or locations (Figure 2-1).

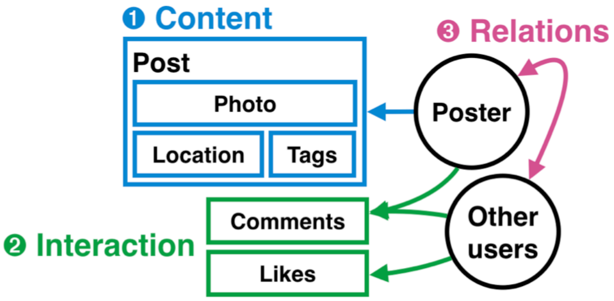

Figure 2. Three perspectives for data analysis. We investigated three aspects of differences between teens and adults and used them to build prediction models.

- *Interaction-based activities* refer to the actions derived from shared content such as Likes or comments (Figure 2-2).
- *Relation-based activities* represent follow-based actions between users (Figure 2-3).

In the following three sections, we present how we measure three activities in detail. Note that many variables used in the analysis are large and non-normalized. To correctly measure the difference in those variables between teens and adults, we used eta-square (i.e., the effect size), denoted as η^2, which refers to the proportion of variance associated with each of the main effects, interactions, and error in an ANOVA study [37]. As a rule of thumb, $\eta^2 = 0.01$, 0.13, 0.26 are considered to be small, medium, and large, respectively.

4.1 Content-based Activities

We first measured how many tags and locations were added to each photo. We then identified the topics in the photos from the tags and checked if the two age groups showed different topic distributions.

4.1.1 Difference in tags, selfies, and locations

Table 2 summarizes the number of tags and selfies (i.e., photos with #selfies and/or #me) per photo and the location diversity of photos. Teens showed higher activities in both tags and selfies per photo. Having more tags per photo often improves the accessibility of the photo, because most social networking sites provide a list of photos with the same tag, when the tag is clicked [28]. Hence, teens' higher tagging activities may indicate their intention to be searched more by other users. From selfies, we can assume that teens showed higher self-representation through photo activities.

Table 2. # Tags and # Selfies per photo, and photo location diversity (T: Teens, A: Adults; *p<0.001; median). Blue cells and orange cells indicate higher influence from teen and adults, respectively.

	P-data			T-data		
	T	A	η^2	T	A	η^2
# Tags / # Photos	1.75	1.70	0.00	3.52	3.35	0.01*
# Selfies / # Photos	0.06	0.02	0.02*	0.07	0.02	0.04*
Location diversity	0.92	2.97	0.06*	1.89	2.48	0.02*

Regarding location diversity, because not all users indicated geo-location in their photos, we randomly chose 1K teens and 1K adults to have the same number of users for the analysis. Given the photos for each user, we first identified the center location that has the minimum distance with all other locations. Then, from the center location, we grouped other locations within a radius of 10 miles. In this sense, the number of groups indicates location diversity. Table 2 shows that adults have photos with higher location diversity than teens. Perhaps this is because adults may have more chances to go to or visit different places (which also relates to the finding in Table 3, where adults have more location tags) and have a more flexible and/or dynamic lifestyle than teens.

4.1.2 Difference in topic distribution and relationship

We also looked into a difference in "content" that people in two age groups create or share. However, analyzing photos (e.g., recognizing objects) using current computer vision techniques is still found to be challenging. One way to infer photo content is looking at the tags in the photos, because research indicates that people tend to add tags that represent the photos they post [13]. Hence, we analyzed the tags in the photos as "proxies" by treating them as a bag-of-words in LDA [4]. We identified 11 latent topics and applied the same method to our datasets and clustered photos into 11 topics (Table 3).

As shown in Table 3, we calculated the percentage of topics from posted photos for each topic group. As a result, we found that both P-data and T-data exhibited similar trends in topics with a particular difference. That is, teens in both P-data and T-data showed more cases in "Mood/emotion" and "Follow/like," while adults showed more cases in "Location" and "Nature." This result is not only consistent with, but also extends what was reported in our prior study [16] using more expanded heterogeneous datasets.

Table 3. Topic distributions (T: Teens, A: Adults; *p<0.001; median). Orange cells (or blue cells) highlight topics that teens (or adults) have more interests with $\eta^2 > 0.01$ from both P- and T-data.

Topics	P-data			T-data		
	T	A	η^2	T	A	η^2
Arts/photo/design	0.10	0.09	0.00*	0.02	0.03	0.00*
Entertainment	0.05	0.04	0.00*	0.08	0.09	0.00
Fashion/beauty	0.08	0.10	0.00*	0.18	0.05	0.12*
Follow/like (T)	0.07	0.04	0.02*	0.07	0.04	0.02*
Food	0.03	0.04	0.01*	0.04	0.06	0.03*
Instagram-tag	0.18	0.18	0.00	0.10	0.09	0.00*
Location (A)	0.12	0.17	0.03*	0.15	0.22	0.03*
Mood/emotion (T)	0.19	0.10	0.08*	0.20	0.14	0.03*
Nature (A)	0.11	0.16	0.02*	0.03	0.05	0.02*
Social/people	0.02	0.03	0.00*	0.10	0.17	0.06*
Sports/wellness	0.05	0.05	0.00*	0.03	0.06	0.02*

4.1.3 Difference in topic entropy

With the eleven topics identified, we defined *topic entropy* as the degree of uncertainty of each user's topic distribution. The entropy of a user's topic distribution X is:

$$Entropy(X) = -\sum_{i=1}^{11} P(x_i)\log_2 P(x_i)$$

where $P(x_i)$ is the probability (i.e., proportion) of topic x_i.

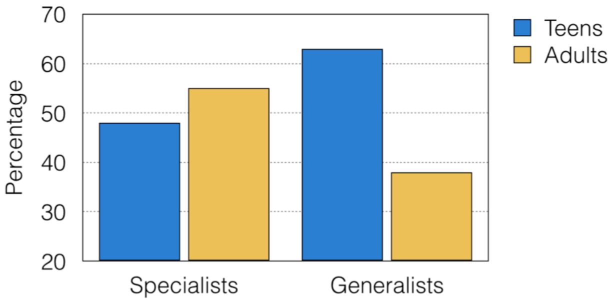

Figure 3. Percentage of specialists and generalists in teens and adults.

A user of high entropy tends to post photos with diverse topics, while a user of low entropy tends to post photos with specific (or fewer) topics. We found that the topic entropy ranges from 0.0 to 3.46 — i.e., a user has minimally 1 topic (2^0=1) and maximally 11 ($2^{3.46}$≈11). With this range, we defined the users with entropy ≤ 1.59 (i.e., ≤ 3 topics) as *specialists* and those with entropy ≥ 3.17 (i.e., ≥ 9 topics) as *generalists* (same topic range).

Figure 3 shows the percentage of specialists and generalists in teens and adults, demonstrating a significant difference (p<0.001). There are fewer specialists but more generalists from teens than adults, meaning that, from the same number of users, more teens are likely to have photos with diverse topics, whereas more adults are likely to have photos with specific topic(s).

4.2 Interaction-based Activities

We now switch our focus to interaction-based activities, which account for Liking and commenting. We measured the differences between teens and adults in the number of Likes, comments, and self-comments and in response time to others' comments.

4.2.1 Difference in Likes, comments, and popularity

Table 4. Number of Likes and comments per photo, number of comments from others, and popularity (T: Teens, A: Adults; *p<0.001; median). Blue cells indicate higher influence from teens.

	P-data			T-data		
	T	A	η^2	T	A	η^2
# Likes / # Photos	38.2	31.8	0.00	33.0	15.0	0.01*
# Comments / # Photos	1.92	1.81	0.01*	1.66	1.40	0.00
# Comments received / # Photos	1.28	1.23	0.01*	1.28	1.12	0.01*
Popularity	10.6	10.4	0.01*	10.0	9.23	0.02*

As shown in Table 4, teens tend to have more Likes, all comments (including self-comments), and comments from others per photo, presenting higher overall interactions than adults. As one's popularly can be inferred by the number of Likes and comments received from others, we devised a measure of *popularity* of a user as follows:

$$Popularity = Log\left(\left(\frac{\#Likes}{\#Photos}\right)\times\left(\frac{\#Comments\_received}{\#Photos}\right)\times\#Followers\right)$$

That is, a user is considered “popular,” if her photos receive many Likes and comments, and she has many followers. The popularity result in Table 4 shows that teens tend to have a higher popularity than adults in both datasets.

In addition, we measured the relationship between entropy-based user groups (i.e., generalists and specialists) and popularity to see an additional difference between teens and adults. Our hypothesis was that specialists would be more popular, because they would have photos with high quality and more focused topics that might attract more other users.

As a result, we found that teens showed nothing significant, but adults showed a strong negative relationship (t(2148)=-8.22, p<0.001). In other words, for adults, the more popular they are, the less photo topics they have. This indicates that, for adults, receiving Likes, comments, and having followers are somewhat related to their photos. For teens, on the other hand, whether they are specialists or generalists does not seem to strongly affect their popularity, perhaps because getting attention in social media is simply a goal for them, and the photo iteself might be less influential.

4.2.2 Difference in self-comments and response time

We examined the number of self-comments (i.e., comments to one’s photos) and the response time to others’ comments. We expected that teens would show more self-comments and shorter response time than adults. As shown in Table 5, this turned out to be true from both datasets.

Table 5. Number of self comments per photo and one’s response time to others comments (minutes) (T: Teens, A: Adults; *p<0.001; median). Blue cells indicate higher influence from teens.

	P-data			T-data		
	T	**A**	**η^2**	**T**	**A**	**η^2**
# Self comments / # Comments	0.30	0.23	0.02*	0.20	0.17	0.02*
Response time to others’ comments	24.6	54.1	0.01*	14.6	30.2	0.00

Regarding response time, we considered a scenario where an original photo poster, *@robinson*, checked one comment (e.g., “Nice pic, where did you take it?”) added to his photo by another user, *@johndoe*. Then, *@robinson* added a new comment (e.g., “*@johndoe*, I took this photo when I visited New York”) and mentioned *@johndoe* in his comment.

Adding user’s name right after the “@” symbol (i.e., mention) has been widely used in social media for replying to another users as well as communicating with others. Based on the scenario above, we measured how quickly posters replied to others’ comments on their photos. As shown in Table 5, teens replied in around 24.6 (P-data) and 14.6 minutes (T-data). The time gap from adults is larger, showing 54.1 (P-data) and 30.2 minutes (T-data). Overall, teens are more engaged in comment-based communications through self-comments and quicker responses.

4.3 Relation-based Activities

Lastly, for relation-based activities, we used the number of followers (i.e., people who follow me) and that of follows (i.e., people whom I follow). Table 6 shows the results.

Table 6. Number of follows and followings (T: Teens, A: Adults; *p<0.001; median). Orange cells indicate higher influence from adults.

	P-data			T-data		
	T	**A**	**η^2**	**T**	**A**	**η^2**
# Followers	464	487	0.01*	285	300	0.01*
# Follows	359	359	0.00	263	328	0.01*

We found that adults tend to have more followers in both datasets and have more follows in T-data. We expected that teens might have more relation-based activities than adults, as they are generally more active in social media, but it turned out to be the opposite. Although we could not draw specific reasons for this from our datasets, we assume that adults might have more opportunities to meet people in real life, having diverse networks, compared to teens whose network might be mostly limited to their friends at school.

4.4 Modeling, prediction, and classification

In this section, we investigate if it is possible to exploit the differences in a number of activities (i.e., features) to build accurate prediction models to classify teens and adults.

We applied four widely adopted supervised learning models — Logistic Regression (LR), Support Vector Machine (SVM) with radial-basis kernel, Random Forest (RF), and Adaptive boosted Decision Trees (ADT). We evaluated all of them using 10-fold cross validation. We preprocessed the features by taking logarithm except topic related features. This in general reduces the side effect of a long-tail distribution. We used a uniform topic distribution for a missing topic distribution when there was no tag.

Table 7. Classifiers performance for content- (C), interaction- (I), and relation-based activities (R) from the merged datasets (P-data and T-data) and through 10-fold cross validation.

Type	Classifier	Accuracy	Precision	Recall	AUC
C	**LR**	0.686	0.661	0.765	0.749
	SVM	**0.726**	0.713	0.756	**0.794**
	RF	0.721	0.722	0.721	0.784
	ADT	0.719	0.699	0.774	0.796
I	**LR**	0.688	0.681	0.710	0.734
	SVM	**0.721**	0.715	0.740	**0.781**
	RF	0.703	0.694	0.730	0.765
	ADT	0.715	0.715	0.718	0.776
R	**LR**	0.539	0.537	0.577	0.534
	SVM	**0.558**	0.567	0.534	**0.567**
	RF	0.525	0.525	0.588	0.520
	ADT	0.555	0.553	0.591	0.567

4.4.1 Baseline

To validate how well our prediction models work, we used two baseline methods to classify a user based on the estimated age: (1) using the user’s profile image; and (2) as the median of user’s

Table 8. Classifiers' performance for P-data, T-data, and P+T data. The highest accuracy is achieved using SVM across all cases.

Data	Classifier	Accuracy	Precision	Recall	AUC
P-data	**LR**	0.720	0.705	0.757	0.783
	SVM	**0.752**	0.740	0.780	**0.826**
	RF	0.748	0.740	0.775	0.825
	ADT	0.748	0.740	0.775	0.825
	Baseline	0.680	N/A	N/A	N/A
T-data	**LR**	0.803	0.826	0.766	0.875
	SVM	**0.824**	0.853	0.784	**0.892**
	RF	0.808	0.837	0.769	0.885
	ADT	0.815	0.836	0.790	0.888
	Baseline	0.640	N/A	N/A	N/A
P+T data	**LR**	0.733	0.724	0.754	0.803
	SVM	**0.779**	0.784	0.770	**0.855**
	RF	0.773	0.789	0.749	0.854
	ADT	0.771	0.778	0.760	0.851

mutual friends' ages. We again used Face++. Regarding the first method, we only considered users who have their face in their profile photo. The second method is justified by an assumption that one's mutual friends are more likely to be in the same age group based on the notion of *homophily* [1]. We believe that these two methods represent reasonable baselines to determine user's ages in Instagram. As a result, using the first method, we observed 68% and 64% accuracy for P-data and T-data, respectively. Using the second method, we observed 62% accuracy for both P-data and T-data.

4.4.2 Classifications by activities

In Table 7, we compared the accuracy of prediction models from each of three perspectives (i.e., content-, interaction-, and relation-based activities) using a merged dataset (P+T data). While accuracy varies depending on the choice of a method and features, we found that (1) both content- and interaction-based features yield 72% accuracy at best and (2) relation-based features perform worse than two baselines do. It does not seem effective to solely look at the number of followers and follows in determining age groups.

4.4.3 Classifications by datasets

Table 8 shows the performance of classifiers trained from each of P-data, T-data, and P+T data. All four classifiers performed similarly on each dataset. The highest accuracy is achieved using SVM with T-data (0.824). All classifiers performed better with T-data than P-data. We conjecture that as all users in T-data have age indicated more directly via tags, the quality of estimated ages in T-data may be better than that in P-data, resulting in more homogenous age group samples (and in turn more effective features).

We also investigated how much each variable has an effect on classification. Table 9 shows coefficients obtained from logistic regression models for a subset of variables. In general, variables with positive (or negative) coefficient imply that teens have higher (or lower) average on those variables than adults have.

Table 9. Coefficients of a subset of variables, content-based (C), interaction-based (I), and relation-based (R), and their statistical results (*p<0.001, **p<0.01, *p<0.05, $^{+}$p<0.01). Variables were log transformed. Blue cells indicate higher influence from teens, and orange cells indicate higher influence from adults.**

Type	Variable	P-data	T-data
C	**# Tags / # Photos**	0.00	-0.66***
	# Selfies / # Photos	1.84***	1.41*
	Entropy	1.13	-3.86***
I	**# Likes / # Photos**	0.79^{+}	4.12***
	# Comments / # Photos	-2.01***	-0.59*
	# Self comments / # Comments	2.47***	0.27
	Response time to comments	-0.11***	-0.13***
R	**# Followers**	-0.49**	-1.25***
	# Follows	0.48***	0.09

We can see that most variables (except the number of tags per photo and entropy) showed similar contributions to classifying teens and adults for both datasets. We should be careful not to interpret this result by individual variables, because such estimates are based on the interaction with other variables (i.e., multicollinearity). Yet, it is still valid to examine the significance of each variable. If the variable is associated with low p-value, it presents unique strength to discriminate teens and adults in underlying datasets. When we look at each variable for each activity, for both P- and T-data, the number of selfies per photo (for content-based activities), the number of comments per photo, response time (for interaction-based activities), and the number of followers (for relation-based activities) showed stronger influence (p<0.05) on classifying teens and adults.

4.4.4 Verify models with an independent test dataset

Our last measurement was to verify our models trained by P-data, T-data, and P+T data with new and non-overlapping 3rd test datasets obtained without using users' profiles or tags. This verification process was necessary because it is important to see the robustness of our models that were built from the activities identified in this paper. Based on our experience, we noted that quite a lot of people create their online account name by using the year of birth; for example, lovelife1997, jingkim92, ksruss83 (all fabricated usernames), etc. From our user samples not included in P-data and T-data, therefore, we looked into their usernames and obtained users who specified year in them. We then manually checked if they are teens or adults by visiting their Instagram page. We confirmed that these users are unique samples not found in P-data and T-data, because they did not have a photo of themselves on their profile and did not share tags that indicate their age. As a result, we had a total of 254 new users (i.e., 151 teens and 103 adults) for the test dataset.

Table 10 shows the performance results of classifiers, *trained* by P-data, T-data, and P+T data, and *tested* using the new 254 test users. Note that our models predict two groups quite accurately, which is around 81% in accuracy and 85% in AUC at best. This result strongly indicates that the activity variables used in our study capture two age groups' behavioral characteristics well across independent user datasets collected through different methods.

Table 10. Classifiers' performance using P-data, T-data, and P+T data for "training" and the new 3rd dataset for "testing."

Data	Classifier	Accuracy	Precision	Recall	AUC
P-data	**LR**	**0.811**	0.820	0.874	**0.852**
	SVM	0.756	0.799	0.788	0.829
	RF	0.717	0.734	0.821	0.771
	ADT	0.780	0.784	0.868	0.845
	Baseline	0.680	N/A	N/A	N/A
T-data	**LR**	0.693	0.695	0.861	0.742
	SVM	0.740	0.815	0.728	0.781
	RF	0.740	0.758	0.828	0.783
	ADT	**0.768**	0.824	0.775	**0.787**
	Baseline	0.640	N/A	N/A	N/A
P+T data	**LR**	**0.795**	0.832	0.821	**0.849**
	SVM	0.764	0.814	0.781	0.816
	RF	0.728	0.789	0.742	0.777
	ADT	0.772	0.804	0.815	0.814

5. DISCUSSION

Data that represent people's use and engagement in social media provide ample opportunities to understand how they behave online. With this rationale, we aimed not only to understand behavioral patterns of two special age groups — teens and adults — through comparative analyses of large datasets collected from two different methods but also to create prediction models for identifying teens and adults.

5.1 Behavioral Differences and Classification

We identified users' ages from their profiles and tags and through manual verification. With a total of 20,000 teens and adults, we investigated a number of different behavioral patterns based on content-, interaction-, and relation-based activities. Overall, we found that teens and adults exhibited *different behaviors* across three activities from both P-data and T-data—i.e., overall, relatively stronger differences in content- and interaction-based activities and weaker ones in relation-based ones.

First, regarding content-based activities, teens tend to have more tags and selfies, and have less location diversity on their photos than adults. For photo topics, teens tend to post more photos with "Mood/Emotion" or "Follow/Likes" tags, whereas adults tend to post more photos with "Location" or "Nature" tags. This extends our previous findings with T-data. We also found more generalists from teens and more specialists from adults. It seems that teens' interest in more diverse topics, as generalists, relates to their developmental stage, and topics might become more focused as teens grow up. Second, regarding interaction-based activities, teens have more Likes and self-comments per photo, and tend to reply to others' comments quicker than adults. We lastly found that popular adults are more likely to be specialists, whereas teens showed no strong relationships.

Overall, our study results show teens' different use patterns compared to adults. Although prior research has extensively studied teens' online behaviors [3,5,15], our study methods and results present unique empirical evidence and insights on teens' online engagement. They also substantiate prior findings; that is, social media use has become pervasive in teens' lives, and much of their social activities is formed and developed through interactions in online space.

In addition, we demonstrate the identification of teens and adults based on their activities up to 82% accuracy. We also measured how classifiers are influenced based on three types of activities. Interaction-based activities were found to be the most influential especially with comment-related actions on classification. At the same time, the number of selfies and that of followers showed a unique strength to discriminate teens and adults from two other activities.

Lastly, we confirm the robustness of our models by testing them with a new user dataset. This demonstrates that our models are comprehensive and empirically validates that the activity variables in our study well represent behavioral characteristics of two age groups.

5.2 Study Implications

Our study suggests several implications and opportunities.

First, social networking sites could provide users with a summary of their activities in a text or visualization format. Users can see a summary of their activities (e.g., how many photos that they posted, Likes or comments that they added or received, when they posted photos or comments, what the topics of their photos are, etc.). With these reports, users will be aware of and can reflect on their activities. This will be beneficial for teens in a practical sense, as awareness has been considered as a key aspect that would allow teens to manage their shared content and privacy [19].

Second, social networking sites could use those summary items to recommend topics with similar interests or users who have shown similar behavioral activities. The current design of Instagram offers some recommendation features; however, mostly, they are limited to user's follow-based or location-based recommendation. Thus, we believe that an *activity-based recommendation* would give users a chance to discover similar topics and to interact with people who show similar interests, activities or in the same age group. Especially teens, as active social media users, may find those recommended topics, contents, and users interesting and useful, given that they tend to be more open to new opportunities and have closer social connections with their peers than other age groups [1]. However, a careful design of supporting this new idea should be taken into account, such as allowing users to control a recommendation process and/or their activity visibility from others.

Lastly, extending the second point, social networking sites could control media content, once they identified teens not from their explicit profile information, but from their usage activities, given that teens are exposed to inappropriate content in social media. Like the parental controls included in many digital television services and mobile software, the content in social media can be filtered and controlled after social networking sites identify vulnerable users. Moreover, there is a potential for detecting and automating cases of misuse from them. Social networking sites could collect and analyze teens' usage to provide suggestions on how to use social media properly to teens as well as to their parents or guardians, if any of misuse or wrong activities is detected. This would be a huge win for many teens, parents, societies, and social media communities, and should be one of the goals in research communities as well.

5.3 Limitations and Future Work

One of the limitations in our study is due from sampling. In general, any non-random samples are inevitably biased, and in our case, both datasets were sampled from users who had their profile photos open or hinted their ages in a bio description, or post age-inferable hashtags. Such users might be less concerned about their privacy than those who do not indicate their ages at all. Therefore, our random samples, P-data and T-data, might be from the users who are biased to share their "age" information willingly in Instagram. The additional 3rd dataset used for testing the robustness of our models might be also biased; however, we believe that users' intention to reveal their ages in this test dataset seems less obvious than P-data and T-data. As such, our findings should be carefully interpreted with a limitation. To generalize our findings beyond Instagram, we plan to repeat the study in other photo-sharing sites such as Flickr or Tumblr as well as other social networking sites such as Facebook.

We used tags to infer the latent topics of photos. However, photos contain abundant semantically rich information that is not well captured in the LDA analysis. There have been growing interests and promising results using deep-learning and its tools, such as *Caffe* [18], for detecting and extracting objects and semantics presented in photos. In future work, we plan to use deep-learning based extraction of features to discriminate teens and adults even better.

While our study was based on quantitative analyses, there are many questions remaining that could be better answered through qualitative studies. For example, we are interested in studying what a Like really means to teens, whether/how they manage their Likes, etc. We plan to conduct qualitative studies to answer these questions better in the future.

6. CONCLUSION

This paper contributes to a better understanding of teens' behavioral characteristics through comparative, large data-driven analyses. We used those patterns for the prediction of teens and adults and obtained 82% accuracy, while our baseline methods present 68% accuracy at best. Our models further achieved 81% accuracy from a new and independent test dataset obtained without access to users' profiles or tags. Our datasets are available at: https://goo.gl/LqTYNv

7. ACKNOWLEGEMENTS

This research was in part supported by NSF CNS-1422215, NSF IUSE-1525601, and Samsung GRO 2015 awards.

8. REFERENCES

1. Aiello, L., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2), 9.

2. Bakhshi, S., Shamma, D., & Gilbert, E. (2014). Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram. *Proceedings of the International Conference on Human Factors in Computing Systems,* ACM, 965-974.

3. Birnholtz, J. (2010). Adopt, adapt, abandon: Understanding why some young adults start, and then stop, using instant messaging. *Journal of Computers in Human Behavior*, 26(6), 1427-1433.

4. Blei, D., Ng., A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

5. boyd, D. (2008). Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media*.

6. boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

7. Duggan, M. (2015). Mobile Messaging and Social Media. *Pew Research Center*.

8. Ellison, N., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication*, 12, 1143–1168.

9. Erikson, E. (1980). *Identity and the life cycle* (Vol. 1). WW Norton & Company.

10. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133-3181.

11. Gilbert, E., & Karahalios, K. (2009). Predicting Tie Strength With Social Media. *Proceedings of the International Conference on Human Factors in Computing Systems,* ACM, 211-220.

12. Han, K., Jang, J., & Lee, D. (2015). Exploring Tag-based Like Networks. *Proceedings of the International Conference on Human Factors in Computing Systems,* ACM, 1941-1946.

13. Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1(1), 21-48.

14. Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of Instagram photo content and user types. *Proceedings of the International Conference on Web and Social Media*, AAAI.

15. Ito, M., Horst, H., Bittanti, M., boyd, d., Herr-Stephenson, B., & Lange, P., et al. (2008). Living and learning with new media: Summary of findings from the Digital Youth Project. *MacArthur Foundation Reports on Digital Media and Learning*.

16. Jang, J., Han, K., Shih, P., & Lee, D. (2015). Generation Like: Comparative Characteristics in Instagram. *Proceedings of the International Conference on Human Factors in Computing Systems,* ACM, 4039-4042.

17. Jang, J., Han, K., & Lee, D. (2015). No Reciprocity in "Liking" Photos: Analyzing Like Activities in Instagram. *Proceedings of the International Conference on Hypertext and Hypermedia*, ACM, 273-282.

18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the International Conference on Multimedia*, ACM, 675-678.

19. Jia, H., Wisniewski, P., Xu, H., Rosson, M. B., & Carroll, J.(2015). Risk-taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors. *Proceedings of*

the International Conference on Computer-Supported Cooperative Work and Social Media, ACM, 583-599.

20. Kaplan, A., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59-68.

21. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15).

22. Lenhart, A. (2015). Teens, Social Media & Technology Overview 2015. *Pew Research Center*.

23. Livingstone, S. (2008). Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New Media & Society*, 10(3), 393-411.

24. Madden, M. (2012). Privacy management on social media sites. *Pew Research Center*.

25. Massey Jr., F. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical association*, 46(253), 68-78.

26. McCallum, A. (2002). "*Mallet: A Machine Learning for Language Toolkit*"

27. Muscanell, N., & Guadagno, R. (2011). Make new friends or keep the old: Gender and personality differences in social networking use. *Journal of Computers in Human Behavior*, 28(1), 107-112.

28. Nov, O., Naaman, M., & Ye, C. (2008). What drives content tagging: the case of photos on Flickr. *Proceedings of the International Conference on Human Factors in Computing Systems*, ACM, 1097-1100.

29. Pfeil, U., Arjan, R., & Zaphiris, P. (2009). Age differences in online social networking - A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Journal of Computers in Human Behavior*, 25, 634-654.

30. Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science* 28, 346(6213), 1063-1064.

31. Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the International Conference on Web and Social Media*, AAAI.

32. Ong, E., Ang, R., Ho, J., et al. (2011). Narcissism, extraversion and adolescents' self-presentation on Facebook. *Journal of Personality and Individual Differences*, 50(2), 180-185.

33. Pater, J., Miller, A., & Mynatt, E. (2015). This Digital Life: A Neighborhood-Based Study of Adolescents' Lives Online. *Proceedings of the International Conference on Human Factors in Computing Systems*, ACM, 2305-2314.

34. Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1-6.

35. Quinn, D., Chen, L., & Mulvenna, M. (2011). Does Age Make A Difference In The Behaviour Of Online Social Network Users? *Proceedings of the International Conference on Internet of Things*, IEEE, 266-272.

36. Smith, A. (2013). Smartphone ownership–2013 update. *Pew Research Center*.

37. Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.

38. Valkenburg, P., Peter, J., & Schouten, A. (2006). Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & Behavior*, 9(5), 584-590.

39. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (85), 2579-2605.

40. Weller, K., & Kinder-Kurlanda, K. (2015). Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research? *Proceedings of the International Conference on Web and Social Media*, AAAI.