Red Is Open-Minded, Blue Is Conscientious: Predicting User Traits From Instagram Image Data

Lisa Branz
Technische Hochschule
Georg Simon Ohm
Nuremberg
lisa.branz
@yahoo.de

Patricia Brockmann
Technische Hochschule
Georg Simon Ohm
Nuremberg
patricia.brockmann
@th-nuernberg.de

Annika Hinze
University of Waikato
Hamilton
annika.hinze
@waikato.ac.nz

Abstract

Various studies have addressed the connection between a user's traits and their social media content. This paper explores the relationship between gender, age and Big Five personality traits of 179 university students from Germany and their Instagram images. With regards to both image features and image content, significant differences between genders as well as preferences related to age and personality traits emerged. Gender, age and personality traits are predicted using machine learning classification and regression methods. This work is the first of its kind to focus on data from European Instagram users, as well as to predict age from Instagram image features and content on a fine-grained level.

1 Introduction

The rise of machine learning methods has opened up novel possibilities to extract patterns from social media data that allow researchers to identify and predict user traits based on data features. This work presents an approach towards finding connections between a user's traits and their social media image data as well as predicting a user's traits from their data. Focusing on Instagram data of German university students, we analyze the connection of image features and image content with a user's gender, age and personality traits. While some studies with similar approaches will be used for comparison, some novel insights regarding the connection of image data and user traits as well as their prediction will be presented.

2 Related Work

A meta-analysis published by (Azucar et al., 2018) gives an overview of 28 studies exploring the connection between digital footprints and users' Big Five personality traits. The authors conclude that personality traits can be inferred with high accuracy from information shared by users on social media platforms. However, only four out of 28 studies included take image content into account and only one of the studies focuses on Instagram. Ferwerda et al. provided deeper insights into the area by exploring the correlation between Big Five personality traits and image features (Ferwerda et al., 2015) as well as image content (Ferwerda and Tkalcic, 2018b) and their suitability for predicting a user's personality (Ferwerda and Tkalcic, 2018a). Kim et al. explored the connection between Instagram image color characteristics (Kim and Kim, 2019) as well as image features, image content and emotional expressions in images (Kim and Kim, 2018) with user traits, such as personality or gender. (You et al., 2014) aimed at predicting a user's gender from posting behavior and image content on Pinterest and identified a number of image content categories suitable for classifying a user's gender. (Song et al., 2018) explored correlations between an Instagram user's image content with their gender and age group, and successfully predicted these user traits from image content. However, only two large age groups were distinguished and age group as well as gender were inferred from a user's profile information and could, therefore, not be verified. (Han et al., 2016) successfully predicted a user's age group from user activities on Instagram. Like (Song et al., 2018), the researchers predicted age only on a coarse-grained level and inferred age information from

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 23–28
Barcelona, Spain (Online), December 13, 2020.

a user's Instagram account. At this point there is, to our knowledge, no study predicting an Instagram user's age on a fine-grained level. This requires collecting a user's age directly through a survey in order to obtain the exact age in years. Moreover, all studies mentioned above were conducted using data from the US, South Korea or randomly selected locations. Hitherto there is, to our knowledge, no study specifically focusing on data from European Instagram users.

3 Method

Participants were students of Technische Hochschule Nuremberg, Germany taking an online survey. The user traits assessed included a participant's age, gender and Big Five personality traits openness (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N). Personality traits were assessed using the German version of the Big Five Inventory 2 (BFI-2) (Danner et al., 2016). The data sample consisted of 16,458 images and survey data of 182 participants (100 female, 79 male, 3 non-binary (excluded)) between the ages of 18 and 36 years with a mean age of 23 years (*SD* = 3.29). Instagram-scraper for Python (rarcega, 2017) was used to download all image content from the participants' Instagram profiles. Saturation-related, hue-related and value-related features were extracted from every image, and average values for each user were calculated. Following the approach in (Ferwerda et al., 2015) and (Kim and Kim, 2018), the Pleasure-Arousal-Dominance model of Valdez and Merhabian (Valdez and Mehrabian, 1994) was adopted. The author found specific combinations of brightness and saturation to represent different emotional states, which are calculated as follows: **Pleasure** = .69 Brightness + .22 Saturation, **Arousal** = -.31 Brightness + .60 Saturation, **Dominance** = -.76 Brightness + .32 Saturation.

Image features used for analysis are displayed in Table 1. Using Google Vision API, every image was annotated with multiple labels describing the depicted objects. This yielded a list of 4,537 unique labels, which were classified into 29 categories (Figure 1). The sum of category occurrences for each user across all images was calculated and normalized in order to account for varying numbers of images among participants. An independent sample t-test was performed in order to compare mean values between gender groups for each image feature and image content score. For the user traits age, openness, conscientiousness, extraversion, agreeableness and neuroticism, statistical correlations were calculated with all of the image feature scores and content scores. To ensure comparability with previous research (((Ferwerda et al., 2015), (Ferwerda and Tkalcic, 2018b), (Kim and Kim, 2018), (Song et al., 2018)), the Random Forest approach was chosen for both regression and classification in the prediction step. A 10-fold cross-validation was employed. F1 score, AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) as well as root-mean-square-error (RMSE) were calculated as performance measures for classifier and regressors, respectively.

4 Results

4.1 Statistical Correlations Between User Traits and Social Media Data

Significant gender differences were revealed with regards to image features as well as image content which are displayed in Table 3. Table 2 displays statistical correlations between image features and age as well as personality traits. Table 4 displays statistical correlations between content categories and age as well as personality traits.

Saturation-related features	A. Average saturation B. Saturation variance
Hue-related features	C. % Red pixels D. % Orange pixels E. % Yellow pixels F. % Green pixels G. % Blue pixels H. % Purple pixels I. % Warm pixels J. % Cold pixels
Value-related features	K. Normalized brightness L. Contrast
PAD	M. Pleasure N. Arousal O. Dominance

Table 1: Image features extracted from all images

	O	C	E	A	N	Age		O	C	E	A	N	Age
A	**-0.17**	0.00	-0.08	-0.10	-0.04	0.08	I	**0.25**	**-0.17**	0.01	0.08	0.03	0.11
B	0.10	**-0.15**	0.04	-0.13	-0.05	0.00	J	**-0.25**	**0.17**	-0.01	-0.08	-0.03	-0.11
C	**0.24**	**-0.20**	0.05	-0.09	0.03	-0.07	L	**-0.21**	0.05	-0.02	0.06	-0.07	-0.14
D	-0.03	0.00	-0.03	0.09	0.10	**0.15**	N	**-0.19**	-0.05	-0.05	-0.07	-0.09	0.02
F	-0.07	0.03	**-0.15**	0.04	0.01	-0.07	O	**-0.18**	-0.09	0.00	-0.02	-0.13	-0.05
G	**-0.25**	**0.20**	0.05	-0.07	-0.03	**-0.20**							

Table 2: Spearman's correlation between image features and user traits. Only features with significant correlations are shown. Significant correlations ($p \leq .05$) are shown in boldface.

1. Food and drinks
2. Animals
3. Botanical
4. Body parts
5. Architecture
6. Beauty and care
7. Landscape and nature
8. Colors
9. Art
10. Clothing and accessories
11. Home and interior
12. Fantasy and fiction
13. Sports
14. Music
15. Dance and performance
16. Fabric and material
17. People
18. Weapons
19. Vehicles and transport
20. Tools and machine
21. Electronics
22. Events, holidays and gatherings
23. Human emotions and behavior
24. Leisure and play
25. Business and education
26. Crafts
27. Services
28. Jewelry
29. Humans

Figure 1: Image content catgeories identified from the data sample

Image Feature	Female	Male	Image Feature	Female	Male
% Orange pixels**	**0.317**	0.267	Beauty and care***	**0.018**	0.004
Brightness*	**0.633**	0.589	Landscape and nature**	0.092	**0.120**
Pleasure*	**0.502**	0.474	Clothing and accessories*	**0.030**	0.021
Dominance*	-0.385	**-0.348**	Jewelry*	**1.510**	0.203
Body parts***	**0.061**	0.0265	Sports**	0.015	**0.026**
Architecture**	0.035	**0.051**	Humans***	**0.073**	0.040

$(p \leq .05), **\ (p \leq .01), ***(p \leq .001)$

Table 3: Significant gender differences in regards to image features and content categories

	O	C	E	A	N	Age		O	C	E	A	N	Age
1	0.08	0.06	0.01	-0.04	-0.01	**0.32**	18	0.12	0.07	0.00	0.13	**-0.16**	0.00
2	0.10	-0.05	-0.12	0.09	0.02	**0.16**	21	**0.23**	-0.14	0.01	0.00	-0.01	**0.15**
6	-0.01	0.06	0.02	0.14	**0.16**	**-0.16**	22	**-0.23**	0.01	0.07	-0.09	0.05	**-0.21**
8	**0.16**	0.00	-0.05	0.07	0.04	-0.06	23	-0.12	0.01	**0.22**	0.03	-0.03	**-0.18**
9	**0.28**	**-0.19**	0.00	0.10	-0.06	-0.14	24	**-0.26**	0.08	0.10	-0.02	0.03	**-0.22**
12	**0.26**	**-0.18**	-0.05	-0.05	0.07	0.04	25	**0.21**	-0.05	0.03	-0.01	0.03	0.14
13	**-0.18**	0.07	0.14	**-0.15**	-0.02	-0.04	26	0.10	-0.02	-0.10	0.06	-0.07	**0.15**
16	**0.15**	0.01	-0.04	0.06	-0.03	0.04	27	0.06	0.08	0.05	**-0.22**	-0.04	0.07
17	**-0.15**	0.03	0.10	0.02	-0.01	**-0.17**							

Table 4: Spearman's correlation between content categories and user traits. Only categories with significant correlations are shown. Significant correlations ($p \leq .05$) are shown in boldface.

4.2 Prediction of User Traits

The prediction of gender using all image features and content categories yielded an F1 score of 0.79 as well as an AUC ROC of 0.78. Performance measures for gender are not compared to previous research as gender was, unlike in previous studies, not inferred from Instagram data. The respective RMSE as well as the interval it refers to for age, openness, conscientiousness, extraversion, agreeableness and neuroticism are displayed in Table 5.

User trait	RMSE [Interval]	User trait	RMSE [Interval]
Openness	**0.62** [1,5] (0.62)	Agreeableness	**0.55** [1,5] (0.56)
Conscientiousness	0.67 [1,5] (0.58)	Neuroticism	0.79 [1,5] (0.67)
Extraversion	0.72 [1,5] (0.61)	Age	**2.88** [18,36] (-)

Table 5: RMSE and reference interval for prediction of user traits with results of (Kim and Kim, 2018) in parantheses. Novel results and results equal to or outperforming results from previous comparable research are shown in boldface.

5 Discussion

In regards to the user trait gender, results align well with (Song et al., 2018) and (Kim and Kim, 2018). Women posting more warm-toned, pleasurable images and displaying a preference for fashion and beauty-related topics as well as social scenes as opposed to men whose images exude more dominance and who tend to display a preference for architecture, sports and outdoor scenes are also in line with common expectations and, therefore, emphasize the validity of results. In regards to age, the preference for social scenes in younger users could be rooted in their stronger desire for social validation and feedback (Somerville, 2013) (Chua and Chang, 2016) in order to build social identity and develop relations (Dhir et al., 2016) (Brown, 1999). In regards to personality traits, only some findings of (Kim and Kim, 2018) and (Ferwerda et al., 2015) (Ferwerda and Tkalcic, 2018b) could be reproduced while in some cases findings even contradict each other (e.g. a preference for red in conscientious users (Kim and Kim, 2018) vs. blue in conscientious users (this study as well as (Labrecque and Milne, 2012), (Navarro et al., 2018))) even though a very similar approach was adopted. Reasons for the discrepancies might include the composition of the user sample in regards to location or socio-economic status as well as current posting and editing trends on social media. However, more research in the area based on larger and more diverse data samples is required to obtain a clearer picture. Preferences that emerged in this work as well as other studies include a preference for art-related topics vs. social scenes in open-minded users (Ferwerda and Tkalcic, 2018b) (Marshall et al., 2015) as well as a preference for warmer colors and social scenes in extraverts and cooler colors in introverts (Kim and Kim, 2018) (Robinson, 1975). These findings are in line with attributes ascribed to these personality traits (Wirtz et al., 2014). The trait agreeableness generally shows a statistical correlation with a user's gender (Vecchione et al., 2012) (Weisberg et al., 2011), with women also scoring significantly higher in our sample. In the light of this connection, both the aversion against sports-related images found in this study as well as the preference for fashion found by (Ferwerda and Tkalcic, 2018b) are in line with expectations. The same holds true for the trait neuroticism, explaining the preference for beauty products and aversion against weapons in images in highly neurotic users. Regarding the prediction of gender, this study outperformed (Kim and Kim, 2018) by a large margin. While (Song et al., 2018) did achieve better results, it has to be noted that their study predicted gender as inferred from image content, not a user's verified gender. Regarding the prediction of personality traits, the prediction in this study outperformed (Ferwerda and Tkalcic, 2018a) in regards to all five personality traits. While (Kim and Kim, 2018) achieved better results predicting conscientiousness, extraversion and neuroticism, this study achieved the same or better results for openness and agreeableness.

6 Strengths and Limitations

While other researchers (e.g. (Song et al., 2018)) inferred a user's gender and age group from their profile information, we tested our results against survey data. Our approach holds two main benefits. Firstly, in order to make age groups clearly distinguishable, (Song et al., 2018) only included teenagers (age < 20 years) and adults (age > 30 years). This approach forced the researchers to exclude all users estimated to be in their twenties, which represents a very large group among Instagram users in general as well as in the data sample for our study. Secondly, it is crucial to mention that predicting a user's age from their Instagram profile might not be reliable. Errors can easily be introduced into the ground truth by users not disclosing their real age, by lack of precision in face recognition tools as well as image editing and the common use of filters. Moreover, in our study we are looking at a user's age on a fine-grained level. This way we are accounting for the fact that particularly among teenagers and young adults, an age difference of a few years can have a large impact on preferences, habits and aversions which cannot be captured as precisely when working with age groups. Another strength of this study is that it is, to our knowledge, the first study of its kind to focus on data from European users, opening up more possibilities for comparison of results across cultures and geographic locations. Limitations include the limited age range of participants as well as the data sample mostly consisting of university students. Future studies should also focus on other age groups and socio-economic groups as well as collect sufficient data from participants of non-binary gender identity. Another limitation is the assignment of each image content label to only one category to avoid statistic interaction between categories. While this makes sense from a statistical point of view, many labels could be assigned to more than one category, which might yield more accurate results in regards to the occurrence of a specific category.

7 Outlook

Possible areas of application could be the non-invasive acquisition of personality data for statistics or research when conducting a survey is not feasible as well as the generation of targeted content in areas such as advertising or usability. However, ethical concerns will have to be addressed when inferring user traits from image data, particularly in areas such as health or employment. The question whether users can consent to the extraction of information from their publicly visible image data with machine learning methods largely incomprehensible to humans will have to be discussed. While some overlap between results of studies in the area is emerging, more research on larger and more diverse data samples is still required to obtain robust results on the connection between user traits and social media image data. Based on the results of this work and related studies, image features and content categories that hold the most and least predictive value in regards to user traits should be identified and used to improved the proposed approach. As data for this study was collected shortly before the onset of the COVID-19 pandemic, future research should explore its impact on how user traits manifest in social media content.

References

Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159.

B Bradford Brown. 1999. " you're going out with who?": Peer group influences on adolescent romantic relationships.

Trudy Hui Hui Chua and Leanne Chang. 2016. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior*, 55:190–197.

Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher Soto, and Oliver John. 2016. Die deutsche version des big five inventory 2 (bfi-2). In *Zusammenstellung Sozialwissenschaftlicher Items und Skalen*.

Amandeep Dhir, Ståle Pallesen, Torbjørn Torsheim, and Cecilie Schou Andreassen. 2016. Do age and gender differences exist in selfie-related behaviours? *Computers in Human Behavior*, 63:549–555.

Bruce Ferwerda and Marko Tkalcic. 2018a. Predicting users' personality from instagram pictures: Using visual and/or content features? In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 157–161.

Bruce Ferwerda and Marko Tkalcic. 2018b. You are what you post: What the content of instagram pictures tells about users' personality. In *The 23rd International on Intelligent User Interfaces, March 7-11, Tokyo, Japan.*

Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, pages 7–10. ACM.

Kyungsik Han, Sanghack Lee, Jin Yea Jang, Yong Jung, and Dongwon Lee. 2016. Teens are from mars, adults are from venus: analyzing and predicting age groups with behavioral characteristics in instagram. In *Proceedings of the 8th ACM Conference on Web Science*, pages 35–44.

Yunhwan Kim and Jang Hyun Kim. 2018. Using computer vision techniques on instagram to link users' personalities and genders to the features of their photos: An exploratory study. *Information Processing & Management*, 54(6):1101–1114.

Jang Hyun Kim and Yunhwan Kim. 2019. Instagram user characteristics and the color of their photos: Colorfulness, color diversity, and color harmony. *Information Processing & Management*, 56(4):1494–1505.

Lauren I Labrecque and George R Milne. 2012. Exciting red and competent blue: the importance of color in marketing. *Journal of the Academy of Marketing Science*, 40(5):711–727.

Tara C Marshall, Katharina Lefringhausen, and Nelli Ferenczi. 2015. The big five, self-esteem, and narcissism as predictors of the topics people write about in facebook status updates. *Personality and Individual Differences*, 85:35–40.

Remedios T Navarro, Edelyn A Cadorna, Lourdes P Llanes, and Michael Esta. 2018. Color preferences and personality traits of cas faculty: Gender differences. *UNP Research Journal*, 22:110–127.

rarcega. 2017. instagram-scraper.

Carl Robinson. 1975. Color preference as a function of introversion and extraversion. *Perceptual and Motor Skills*.

Leah H Somerville. 2013. The teenage brain: Sensitivity to social evaluation. *Current directions in psychological science*, 22(2):121–127.

Junho Song, Kyungsik Han, Dongwon Lee, and Sang-Wook Kim. 2018. "is a picture really worth a thousand words?": A case study on classifying user attributes on instagram. *PloS one*, 13(10):e0204938.

Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology: General*, 123(4):394.

Michele Vecchione, Guido Alessandri, Claudio Barbaranelli, and Gianvittorio Caprara. 2012. Gender differences in the big five personality development: A longitudinal investigation from late adolescence to emerging adulthood. *Personality and Individual Differences*, 53(6):740–746.

Yanna J Weisberg, Colin G DeYoung, and Jacob B Hirsh. 2011. Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 2:178.

Markus Antonius Wirtz, Janina Strohmer, and Verlag Hans Huber. 2014. Dorsch-lexikon der psychologie.

Quanzeng You, Sumit Bhatia, Tong Sun, and Jiebo Luo. 2014. The eyes of the beholder: Gender prediction using images posted in online social networks. In *2014 IEEE International Conference on Data Mining Workshop*, pages 1026–1030. IEEE.