'Expected most of the results, but some others...surprised me': Personality inference in Image Tagging Services

Maria Kasinidou [1], Styliani Kleanthous [1,2], and Jahna Otterbacher [1,2]

Cyprus Center for Algorithmic Transparency, Open University of Cyprus [1]
CYENS Centre of Excellence [2]
{maria.kasinidou, styliani.kleanthous,jahna.otterbacher}@ouc.ac.cy

Abstract. Image tagging APIs, offered as Cognitive Services in the movement to democratize AI, have become popular in applications that need to provide a personalized user experience. Developers can easily incorporate these services into their applications; however, little is known concerning their behavior under specific circumstances. We consider how two such services behave when predicting elements of the Big-Five personality traits from users' profile images. We found that personality traits are not equally represented in the APIs' output tags, with tags focusing mostly on Extraversion. The inaccurate personality prediction and the lack of vocabulary for the equal representation of all personality traits, could result in unreliable implicit user modeling, resulting in sub-optimal – or even undesirable – user experience in the application.

Keywords: Algorithmic bias, cognitive services, personality, image analysis

1 Introduction

Image analysis algorithms, with their seamless functionalities, have been a boon to commercial technologies where modeling and applying user characteristics is vital. Tech giants (e.g., Google, Amazon, Microsoft) have all released "Cognitive Services" that are readily available for almost anyone to use through their websites, and for developers to incorporate into the software they are developing using their APIs. Users have indirect interaction with these tools, whether they are recognizing this or not, and they likely take their outputs for granted.

It is important to understand that image tagging APIs do not always treat people images in a fair and predictable way. Recent work [1, 8, 11] demonstrates that these services are anything but socially just, underscoring the need to be critical when using computer vision algorithms that shape human interactions. For instance, there are reports of gender misclassification, particularly for people with darker skin as compared to people with lighter skin, and women compared to men [2]. Black men were more likely to be tagged with negative emotions as compared to white men [11], when using Face++ and Microsoft's Face API.[1]

[1] https://azure.microsoft.com/en-us/services/cognitive-services/face/

Recent work [8] demonstrated that in neutral images, where only one person is depicted, most taggers had very low accuracy on using gender-related tags appropriately. Similarly, for emotion inferences [9], Clarifai uses emotion-related tags to describe two-thirds of the images, far more than Google or Imagga. Only Clarifai uses words that infer a person's traits, with images of men being described more often with trait tags, as compared to those of women and with Asians being described with the fewest trait-related tags.

Given that these are "black box" algorithms, understanding the source(s) of this behavior, is not straight forward. This becomes especially important when image tagging algorithms are used for implicitly modeling users' personalities in social applications and media (e.g., dating apps), a method that is much less intrusive than having the user to fill in a personality questionnaire. Previous work on personality analysis through images, focused mainly on visual and content features of social media images uploaded by users and suggest that these features are reliable enough for predicting personality [3]. However, there is a consensus that some traits are easier to predict than others through photos. Guntuku et al. [6] used Imagga for tagging a number of images posted and liked by Twitter users, to examine whether Big Five personality traits are related to this activity. Results indicate that only Openness and Neuroticism could be predicted.

Currently, we follow a human in the loop audit approach to examine the behaviors of two popular tagging services, Imagga and Clarifai, aiming to understand the extent to which they predict users' personalities, and if they do so accurately. The research questions we address are: RQ1 - Which traits of the Big Five are described in the APIs' tags? RQ2 - How accurate are the personality-related tags? RQ3 - How do users feel about the tags describing their photos?

2 Methodology

In order to answer the above questions we recruited 38 participants (20 women, 18 men), all being undergraduate university students. Participation was voluntary and all participants provided written, informed consent for their data (IPIP scores, photos and the APIs' outputs) to be used. To answer RQ1 - we asked participants to interact with two popular APIs (Imagga, Clarifai), collecting the output tags from each. For uniformity, participants were asked to take a picture (selfie) with a neutral facial expression, without a background. Participants were asked to upload the selfie to the two APIs and collect the generated tags along with their confidence scores. To answer RQ2 - participants were asked to complete the IPIP questionnaire, a 50-item questionnaire that assesses Goldberg's [5] Big Five factors. To answer RQ3 - participants responded in free text to the following three questions. *How did you feel when you saw the tags that described your photo? Which tags do you consider to be the most representative for you? Which tags do you consider to be the least representative for you?*

Identifying Personality Related Tags The most reliable methods for predicting measurable elements of personality are those based on traits. One of the most stable models is the five-factor solution based on adjectives. Such

a model is based on the Big Five traits, for example, Goldberg's International Personality Items Pool (IPIP).[2] Norman's Taxonomy of traits was an initial attempt to create clusters of English language adjectives that could be used to characterize a person under the Big Five traits [10]. Norman published 1.431 trait adjectives under 75 categories representing each Big Five trait. Goldberg [5] expanded on Norman's Taxonomy of trait descriptive adjectives adding 479 synonym adjectives and developed the revised synonym clusters.

For this work, Princeton WordNet[3] was used to further extract synonyms of the trait adjectives based on Goldberg's revised trait synonyms.[4] For each term from Goldberg's sets, we conducted a search in WordNet, extracting the synonyms based on the meaning of the original word. We used this collection of trait adjectives and the synonyms extracted from WordNet to identify personality descriptive tags output by the two APIs. Imagga provided: 20 distinct tags for Extraversion (14 from the taxonomy and 6 from WordNet), 13 for Openness/Intellect (4 from the taxonomy and 9 from WordNet), 7 for Agreeableness (5 from the taxonomy, 2 from WordNet), 3 for Conscientiousness (from the taxonomy) and 2 for Neuroticism/Emotional Stability (from the taxonomy). Clarifai provided 6 distinct tags for Extraversion (4 from the taxonomy and 2 from WordNet), 4 for Openness/Intellect (1 from the taxonomy and 3 from WordNet), 1 tag for Agreeableness (from the taxonomy) and 1 tag for Conscientiousness (from the taxonomy) and no tags for Neuroticism/Emotional Stability.

3 Analysis and Results

A preliminary descriptive analysis showed that the mean across the 38 participants for each of the five traits, as revealed by the IPIP, was as follows: Extraversion - 30.61 (SD = 6.57), Agreeableness - 37.55 (SD = 5.326), Conscientiousness - 37.16 (SD = 6.58), Emotional Stability - 28.26 (SD = 6.97), Intellect - 35.74 (SD = 4.46). The scores on Extraversion, Conscientiousness and Emotional Stability were approximately normally distributed, while those of Intellect and Agreeableness were slightly skewed to the left.

Trait Representation in the Image Tagging APIs' Output (RQ1): We identified a total of 45 tags from Imagga and 12 from Clarifai, that are related to personality.[5] Tags related to Extraversion (26) are more numerous than those relating to other traits, in both APIs. Extraversion is followed by Intellect (17) with more tags than the other three traits. Interestingly, for Emotional Stability (2), Clarifai did not have any related tags and Imagga had only two.

[2] https://ipip.ori.org/

[3] https://wordnet.princeton.edu/

[4] WordNet is the most widely used English lexical database which includes nouns, verbs, adjectives, and adverbs. The words are organized and linked based on their lexical concept (set of synonyms).

[5] It is important to remind the reader that these services are effectively "black boxes," thus, the complete list of their tags is not publicly available, not even to the developers who are incorporating them in the software they are developing.

Table 1. Factor analysis models

Number of factors	χ^2 test statistic, p-value		Cumulative variance explained
7	463.1, 399	0.0146	0.655
8	414.15, 370	0.0563	0.678
9	371.11, 342	0.134	0.706

Imagga has a richer vocabulary of personality-related tags compared to Clarifai, which had only one tag for Conscientiousness (4) and one for Agreeableness (8). Previous work reported that Extraversion was the easiest to predict from images on social media [12], however, to our best knowledge, there is no previous work on representing personality traits by Image Tagging APIs, to which to compare our results.

Personality Prediction Accuracy (RQ2): To this end, we analyze which traits are typically described by the taggers, and how their use correlates to our participants' scores on the IPIP personality questionnaire. First, we observed that while Imagga and Clarifai have several personality-related tags, many are used very sparingly. Therefore, we considered the tags which were used to describe at least five of the 38 selfies. The tags used less than five times were eliminated from the analysis. Thus, the final set analyzed comprised 36 tags. Furthermore, since many of our tags are conceptually similar (e.g., smile, laugh), we subjected the matrix of tag confidence scores, which for Imagga and Clarifai range from 0 to 100, to a factor analysis, to reduce the dimensions of the dataset and to create a new set of explanatory variables (i.e., tag scores) that are orthogonal to each other. This allows us to study the entire "profile" of the person inferred by the taggers, rather than considering the use of individual tags. It also captures the degree of certainty the tagger has about a particular trait adjective.

Factor analysis. We first produced a scree-plot, a diagnostic tool for determining the optimal number of factors to account for the variance in the data [4]. The plot suggested a solution between five and nine factors. We used the factanal function in R,[6] which uses the maximum likelihood approach to fitting a common, orthogonal factor model, with varimax rotation. The function also outputs results from a chi square test, which evaluates the null hypothesis that the model fit is satisfactory. As shown in Table 1, the model with nine factors fits the data well (i.e., the null hypothesis cannot be rejected). The results of an orthogonal rotation of the 9-factor solution are shown in Table 2.

Interpretation of factors. The items heavily loaded onto Factor 1 indicate a casual, mature and good person with a lack of "flashy" characteristics such as sexy, elegant, etc. This factor was labelled, "Easy-going characteristics". Four items loaded onto the second factor, and related to individuals' positivity (e.g., happy, smiling). This factor was labelled, "Outward positivity". The three items loaded onto Factor 3 identify the quality of being competent, such as the quality of having sufficient knowledge or confidence. However, such individuals are not

[6] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/factanal.html

Table 2. Rotated Factor Loadings for the Nine-Factor Solution

	F1	F2	F3	F4	F5	F6	F7	F8	F9
Casual	0.523		0.102	0.331	0.624	0.345	-.173	0.215	
Relaxed	0.282	-.194	0.254	0.834		0.101	-.279		-.122
Friendly	0.101	0.806	0.137			0.228			0.298
Good	0.581	-.584	0.120	0.464	0.222			-.142	
Nice	-.271	0.189		-.106		-.604			
Serious	0.194		0.233	0.270	0.186		-.142	0.494	
Mature	0.649	-.548	0.245	0.273	0.151		0.156	-.153	
Care	-.298	.298	.331		-.468	.661		-.144	0.133
Expression	0.230	-.452					0.284	0.560	0.123
Confident	0.499	0.271	0.640	0.276			0.161	0.274	
Happy	-.224	0.802	-.207		-.206	-.157	0.150	-.270	
Sexy	-.865			-.151	0.145	-.200			
Smiling	0.168	0.786	0.112		-.171		0.117	-.335	-.165
Smile		0.768		-.261	-.102	-.115	0.144	0.120	
Expressive	0.453	-.232	0.159	0.467			0.370		-.269
Cheerful		0.230	-.250					-.108	0.695
Happiness	0.101	0.343	-.830		-.222	0.107	0.185		-.189
Confidence	0.243	0.280	0.596	0.384	-.145	-.125	0.168		
Joy			-.566	-.119		-.209	0.205	-.174	0.172
Sensual	-.861		-.136		-.118	-.206	0.109		-.111
Sensuality	-.835			-.228	-.109				
Cool	0.420	-.386	0.239	0.534	0.220	0.247	-.164	0.107	
Smart	0.343	0.197	0.532	0.482	0.199				
Thoughtful	0.169	-.182	0.140				-.814		
Modern		-.156			0.554		-.169	0.252	0.239
Elegance	-.687		0.166	-.246		0.187		0.105	0.332
Elegant	-.588		0.237	-.175	0.210	-.275			0.117
Fashionable		-.143	0.129	0.133	0.683		0.158		
Trendy	0.242	-.150	0.106	0.798	0.189		0.118	0.213	0.150
Glamor	-.651		-.186		0.107	0.168		-.207	
Prop. Var.	**0.183**	**0.122**	**0.093**	**0.092**	**0.059**	**0.046**	**0.040**	**0.039**	**0.033**

characterized as happy/joyful, as evidenced by the negative loadings on these characteristics. Thus, Factor 3 was labelled, "Competence".

The five items that loaded onto Factor 4 relate to tags that describe a positive impression an individual has on others (e.g., cool, trendy). This factor was labelled, "Positive impression on others". Three items loaded onto Factor 5, which are related to physical characteristics that tend to have a positive impression on others (e.g., modern, fashionable, casual). This was labelled, "Positive impression on others, physical characteristics." Items with heavy loadings on Factor 6 are related to feeling concern or interest in something or someone. This factor was labelled, "Caring". Factor 7 has lots of negative weights, in particular "not thoughtful". This factor was labelled, "Less positive." Items for Factor 8 related to being serious and not happy. This factor was labelled, "Serious demeanor."

Table 3. Linear regression to predict the personality traits using the nine factors.

	E	C	ES	Log (I +1)	Log(A + 1)
(intercept)	30.61***	37.16***	28.26***	3.60***	3.64***
F1	-0.67	-1.14	0.52	0.03	-0.02
F2	-0.40	1.56	-0.33	-0.01	0.02
F3	0.56	0.49	2.03	-0.01	0.003
F4	-1.06	-0.58	0.35	0.02	-0.03
F5	-0.20	-0.95	1.31	0.02	0.03
F6	0.84	-0.88	0.70	0.03	-0.02
F7	1.45	-0.60	-0.21	0.03	-0.03
F8	-2.47*	-2.24	-0.09	0.001	-0.05*
F9	-.58	0.44	2.27	-0.002	0.005
R-squared	0.24	0.25	0.23	0.20	0.27

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Finally, Factor 9 had strong loadings on cheerful, elegant and stylish. This factor was labelled, "Cheerful".

To study the relationship between each of the nine factors - representing how the taggers perceived the depicted people - and the personality scores of the participants, we used linear regression. In particular, we regressed the scores that followed a normal distribution (Conscientiousness, Emotional Stability, Extraversion) onto the nine factors. We took the log transformation on Intellect and Agreeableness, which was approximately normal. Table 3 presents the regression model in which the nine factors are used to predict each personality trait.

Factor 8, which represents a serious expression, has a significant negative correlation to Extraversion and Agreeableness, which is quite sensible. However, for the other traits, none of the factors shows a significant correlation. It is interesting to note that the factors that are more important in explaining the variance in the way people are described (F1, F2), each of which explains more than 10% of the variance, do not appear correlated to personality traits.

Participants' view on Image Tagging APIs' Output (RQ3): Thematic analysis [7] applied, by two researchers, to participants' free text responses to uncover the key factors participants considered when they saw the tags assigned to their selfies by the APIs. Five categories, which were not mutually exclusive, have emerged. The number in the parentheses indicates the frequency with which participants' responses mentioned that concept.

Accuracy (19). The participant considered the tags as true or correct.

Weird (9). The participant "felt weird" about the tags.

Enthusiasm/Surprised/Impressed (13). Participants felt surprised because the tags were unexpected.

Expected/Unexpected (11). Whether or not the participant was expecting the output tags.

Subjectivity (1). subjective tags and based on real characteristics.

As observed, almost half of the participants discussed the tags' accuracy (either accurate or inaccurate). Nine responses mentioned that they felt weird

Table 4. Most (left) and least (right) representative tags by participants gender .

	Men	Women		Men	Women
Personality Related (n=34)	9	25	Personality Related (n=32)	16	16
Age Related (n=21)	7	14	Age Related (n=12)	8	4
Gender Related (n=35)	16	19	Gender Related (n=16)	6	10
Appearance Related (n=49)	21	28	Appearance Related (n=23)	8	15
Other (n=17)	12	5	Other (n=26)	8	18

about how an image tagging algorithm could recognize so many things from an image with no expression and no background or because they could not understand how the API chose specific tags. Interestingly, only one participant directly mentioned the subjectivity of the tags.

Participants were also asked to discuss the tags they deemed as the most/least representative ones. To reveal the key themes, we followed the same approach as above. Five themes emerged, Table 4 presents the number of responses in which participants mentioned each of the five categories as the most representative tags, broken out by participant gender. Both men and women mentioned that the appearance-related tags were the most representative for them (e.g., brunette, pretty). This resonates with previous findings on users' tendency to present themselves in a socially desirable way [3]. As observed, women expressed more often than men the feeling that personality-related tags were the most representative ones. Another interesting finding is that all personality-related tags mentioned were positive tags (e.g., happiness, friendly) or negative tags (e.g., anger, fear). Both men and women mentioned - although less often than other types - tags related to their age. Finally, men and women mentioned gender-related tags as the most representative with similar frequencies.

As can be seen, personality-related tags were often mentioned by both men an women. In particular, three respondents mentioned in their answers that all the tags related to emotions were the least representative (e.g. *"I think the tags related to the feelings were less representative of me."* - P1).

4 Concluding Remarks

Recent work has highlighted several concerns with respect to the use of image analysis algorithms in processing people images. In particular, previous studies demonstrated that computer vision algorithms are often producing less accurate results for some groups of depicted persons over others. The present work has echoed the concerns but from a different perspective. We found that while image tagging algorithms often output personality-related tags when processing a selfie, the tags do not correlate to the depicted person's actual personality.

This is clearly not an objective task especially when is used to implicitly build a user model. This finding was very much confirmed in the qualitative responses of participants, who overwhelmingly felt that tags related to their physical appearance were the most representative compared to personality related tags. In

conclusion, developers should be aware of the behaviour of the image tagging algorithms when using them for implicitly modeling users' personalities.

5 Acknowledgments

This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under agreements No. 739578 (RISE) and 810105 (CyCAT).

References

1. Barlas, P., Kleanthous, S., Kyriakou, K., Otterbacher, J.: Social b(eye)as in image tagging algorithms: Human and machine descriptions of people images. Proceedings of AAAI ICWSM **13**(01), 583–591 (2019)
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR, New York, NY, USA (2018)
3. Celli, F., Bruni, E., Lepri, B.: Automatic personality and interaction style recognition from facebook profile pictures. In: Proceedings of the 22nd ACM International Conference on Multimedia. p. 1101–1104. ACM, New York, NY, USA (2014)
4. Everitt, B., Hothorn, T.: An Introduction to Applied Multivariate Analysis with R (UseR!) (01 2011). https://doi.org/10.1007/978-1-4419-9650-3
5. Goldberg, L.: An alternative "description of personality": The big-five factor structure. Journal of personality and social psychology **59**, 1216–29 (01 1991)
6. Guntuku, S.C., Lin, W., Carpenter, J., Ng, W.K., Ungar, L.H., Preoţiuc-Pietro, D.: Studying personality through the content of posted and liked images on twitter. In: Proceedings of the 2017 ACM on Web Science Conference. pp. 223–227. WebSci '17, ACM, New York, NY, USA (2017), http://doi.acm.org/10.1145/3091478.3091522
7. Herring, S.C.: Web content analysis: Expanding the paradigm. In: International handbook of Internet research, pp. 233–249. Springer (2009)
8. Kyriakou, K., Barlas, P., Kleanthous, S., Otterbacher, J.: Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In: Proceedings of AAAI ICWSM. vol. 13, pp. 313–322. AAAI, California, United States (2019)
9. Kyriakou, K., Kleanthous, S., Otterbacher, J., Papadopoulos, G.A.: Emotion-based stereotypes in image analysis services. In: Adjunct Publication of the 28th ACM UMAP Conference. pp. 252–259 (2020)
10. Norman, W.T.: 2800 personality trait descriptors: normative operating characteristics for a university population. Uni. of Michigan, Ann Arbor, Michigan (1967)
11. Rhue, L.: Racial influence on automated perceptions of emotions. Available at SSRN 3281765 (2018)
12. Silveira Jacques Junior, J.C., Güçlütürk, Y., Perez, M., Güçlü, U., Andujar, C., Baró, X., Escalante, H.J., Guyon, I., Van Gerven, M.A.J., Van Lier, R., Escalera, S.: First impressions: A survey on vision-based apparent personality trait analysis. IEEE Transactions on Affective Computing pp. 1–1 (2019)