

A deep multimodal learning approach to perceive basic needs of humans from Instagram profile

Journal:	<i>Transactions on Affective Computing</i>
Manuscript ID	TAFFC-2020-12-0356.R1
Manuscript Type:	Regular
Keywords:	Social media, Multi-modal learning, Multi-label classifier, Choice theory, Deep learning, Bag of Content

SCHOLARONE™
Manuscripts

A deep multimodal learning approach to perceive basic needs of humans from Instagram profile

Mohammad Mahdi Dehshibi¹, Bitu Baiani², Gerard Pons³, David Masip¹

¹Department of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain

²Department of Psychology, Islamic Azad University, Science and Research Branch, Tehran, Iran

³Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands

Nowadays, a significant part of our time is spent sharing multimodal data on social media sites such as Instagram, Facebook and Twitter. The particular way in which users present themselves to social media can provide useful insights into their behaviours, personalities, perspectives, motives and needs. This paper proposes to use multimodal data collected from Instagram accounts to predict the five basic prototypical needs described in Glasser’s choice theory (i.e., *Survival, Power, Freedom, Belonging, and Fun*). We automate the identification of the unconsciously perceived needs from Instagram profiles by using both visual and textual contents. The proposed approach aggregates the visual and textual features extracted using deep learning and constructs a homogeneous representation for each profile through the proposed *Bag-of-Content*. Finally, we perform multi-label classification on the fusion of both modalities. We validate our proposal on a large database, consensually annotated by two expert psychologists, with more than 30,000 images, captions and comments. Experiments show promising accuracy and complementary information between visual and textual cues.

Index Terms—Social media, Multi-modal learning, Multi-label classifier, Choice theory, Deep learning, Bag of Content.

I. INTRODUCTION

MENTAL health is an integral and essential component of health affecting not only individual attributes but also social, cultural, environmental, political, and economic factors. Therefore, preventing them at an early stage can result in substantial cost and life savings for societies [1].

In our technology-connected communities, clues about people’s behaviours, emotions, and psychological conditions can be found in their use of social media as to what they read, like, post, and follow [2, 3]. These behavioural patterns have motivated researchers in the field of psychology, natural language processing, and affective computing to introduce new solutions and approaches for spotting early warning signs of mental health issues. Their findings arguably improve the preventive measures’ processes, from detection to better assessing treatments once a person with mental disorders has been identified [4, 5, 6, 7].

Although the use of social media is not limited to English-speaking users, the non-English parts are relatively unexplored by their own perceptions of mental health and social stigma. For instance, Ramirez-Esparza et al. [8] reported that depressed users who have written in Spanish are more likely to mention relationship problems than depressed users who have written in English. Cuijpers et al. [9] have shown the importance of online interventions to engage with people from different ethnic backgrounds who have suffered from depression and anxiety.

The study of social media data belonging to non-English users could help to build inclusive and diverse tools and models for addressing mental health issues in people with diverse cultural or linguistic backgrounds. However, all psychological theories, schemes and questionnaires must be culturally

adapted¹ by expert psychologists so as to be used in assessing non-English-speaking social media users. In this context, the use of the visual medium may be a meaningful strategy for conquering linguistic barriers. Images often evoke common concepts [11]. Even if a particular image evokes a different concept for a person, the impact of this exception can fade to a larger scale. For instance, in Beck’s cognitive depression theory, affected individuals tend to perceive themselves in mostly negative and dark environments [12].

The choice theory (also known as reality therapy) developed by Glasser [13] expresses that our behaviour is driven by one or more of our basic needs, and these basic needs drive our choices. Needs that if not met, could be the source of a lot of personal unhappiness, and as a consequence, cause the onset of mental health problems [14, 15, 16]. Belonging, Power, Freedom, Fun, and Survival are these basic needs which characterised by a Personal Picture Album (PPA). PPA is a specific place to store mental images of people, places, things, values and beliefs that are important to us and can satisfy at least one or more of our basic needs [17]. Social media platforms like Instagram enable their users to reflect their mental images through what they read, like, post and follow. This kind of picturing mental images not only aligns with three out of 10 axioms of Glasser’s choice theory² but has also been used in studies that target social media self-disclosure behaviour [18, 19, 20].

This study used the choice theory to categorise users’ profiles considering five basic needs from a broader yet language-

¹Cultural adaptation is defined as “the systematic modification of an evidence-based treatment or intervention protocol to consider language, culture, and context in such a way that it is compatible with the individual’s cultural patterns, meanings, and values” [10].

²[A2] All we can give another person is information. [A6] We can only satisfy our needs by satisfying the pictures in our Quality World. [A7] All we do is behave.

Manuscript received XYZ; revised XYZ. Corresponding author: M. M. Dehshibi (email: mdehshibi@uoc.edu)

independent perspective. To this end, we have studied how people implicitly contribute to unmet basic needs by *choosing* content to share on their Instagram profile in order to draw a personal picture of their quality-world. We also explored how visual and textual contents can tell us about (1) those whom we care about and who care about us (Belonging); (2) those whom we respect and who respect us (Power); (3) those who allow us to think for ourselves and make choices (Freedom); (4) those with whom we laugh (Fun); and (5) those who provide us with the conditions for physical and emotional security (Survival). We proposed a multimodal and multi-label deep learning approach that perceives the five basic needs of users from their Instagram profiles.

In two sessions, with an interval of 1.5 years, we collected data from 86 public Instagram profiles (excluding business, influencers and celebrities) by observing Instagram³ and Facebook⁴ data policies. The owners of these profiles were native Persian and Spanish speakers living in Iran and Spain during data collection. Out of 86 profiles, owners of 10 profiles met with the lead psychologist for at least 15 therapy sessions and consented to the use of their data for this study. *Two psychologists visited only the visual contents of profiles and for each profile a consensus ground-truth based on Glasser's choice theory was then given.*

In the proposed architecture, however, we use both visual and textual modalities to build a multimodal and multi-label classifier. To extract visual features, we use the Places-CNN and YOLO object detector [21] with a modified detection sub-network that has been trained with the Places2 [22] and the Microsoft COCO datasets [23], respectively. To represent textual content, we fine-tune the FastText embedding model [24] with all the words in English, Spanish, Persian and Turkish that appear in our dataset. To integrate the outputs of these three streams, we propose *Bag-of-Content* (BoC). The aggregated features are then passed to Multi-Label Learning with GLOBAL and loCAL Label Correlation classifier [25] to perceive the five basic needs. We evaluate the proposed architecture with our dataset, which reveals promising results.

Indeed, our contribution is twofold: (1) we introduce a new dimension of research in the field of affective computing by undertaking an exploratory and interdisciplinary study of the automatic prediction of five basic needs in accordance with Glasser's choice theory; (2) rather than just providing a benchmark for this new dimension [26], from a technical perspective, we propose *Bag-of-Content* (BoC) to fuse and minimise the dimensionality of the multimodal features. Experimental results indicate improved accuracy using the proposed BoC approach.

The rest of this paper is organised as follows: Section II surveys previous studies. Section III details data gathering, including the sampling, statistics, ethical concerns, and labelling. Section IV describes the proposed deep approach to extract visual and textual features, Bag-of-Content based features fusion, and multi-label classification. Section V presents experiments results. Finally, Section VI concludes the paper.

II. RELATED WORK

The sharing of multimodal data (e.g. image, video, text) has become an essential part of online social experience. Studies have found that these data can help to detect early warning signs of changes in physical and mental health, personality, and users needs [4, 6, 7, 9].

Reece et al. [4] proposed a computational model to predict depression signs in users' Instagram data and showed that depression indicators are effectively identifiable within six months before diagnosing the trauma by health professionals. This progress, compared to the average 19-month delay between trauma onset and diagnosis experienced by the individuals, can provide a framework for an accessible, accurate and cost-effective screening of depression, where in-person assessments are difficult or costly.

Kircaburun and Griffiths [27] asked 752 university students to complete a self-reported survey, including Instagram addiction and self-liking scales. Results revealed that agreeableness, conscientiousness, and self-liking are negatively associated with Instagram addiction, while daily Internet usage is positively associated with Instagram addiction. Nevertheless, the lack of providing methodological details for the assessment of users restricts the possibility of making effective use of the findings of this analysis.

Kim and Kim [26] utilised computer vision methods to find the relationship between photos posted by users on Instagram and personality traits already assessed by an online survey. Content categorisation was done by counting the number of faces, analysing the facial expression, and pixel-derived features using the Microsoft Azure Computer Vision API [28]. However, since they believed that expressing oneself by photo is more straightforward than by providing text, they did not examine textual contents that appeared in biography, captions, and comments.

Pampouchidou et al. [29] surveyed automatic depression assessment studies in which the visual cues were used. They addressed several research questions, including the number of modalities, facial signs, experimental protocols for data acquisition, feature descriptors, decision-making processes, and evaluation metrics. They concluded that results are consistent with the social withdrawal, emotional-context insensitivity, reduced reactivity hypotheses of depression, the gender dimension and the significance of complex features/multimodal approaches through quantitative study. Similarly, they argued that to achieve clinically useful results, visual cues need to be supplemented by information from other modalities.

Surveying recent studies implies that predictive methods are not mature enough to detect mental disorders effectively. Limitations include: (1) systematic gaps in clinical research questions to distinguish between different disorder sub-types; (2) inappropriate and inadequate generalisation of predictive methods due to targeting only one class of mental health problems, e.g., depression or addiction; (3) linguistic bias to English speaking users to alleviate the difficulty of adapting psychological theories to other cultural structures; (4) inability to build a representative latent space when a particular modality is used. In this study, we used the choice theory to

³<https://help.instagram.com/519522125107875>

⁴<https://www.facebook.com/help/203805466323736>

assess mental health from a broader yet language-independent perspective [15, 16]. We took advantage of both visual and textual modalities to explore the relationship between the five basic needs and the corresponding latent space to a user's Instagram profile.

III. DATA COLLECTION

In this research, we collected the visual and textual contents of 86 Instagram profiles (10 private and 76 public profiles) in two phases between January 2019 and September 2020 using the Instagram Application Programming Interface. Each profile contains images and a JSON file. The JSON file contains biography (known as bio), feed caption, comments, and geotags. The textual content is in English, Spanish, Turkish, and Persian, including hashtags and emojis. **In total, we collected 30,080 feeds (each feed may have multiple images) in the first phase and 7,450 feeds in the second phase.** Figure 1 shows an Instagram profile.

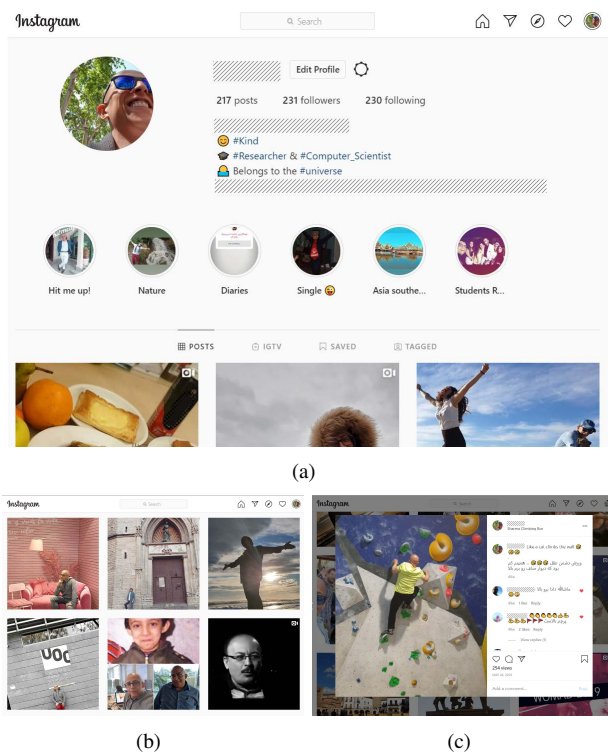


Fig. 1. A sample Instagram profile with (a) bio, (b) feeds and (c) post, including caption and comments. The textual content of this sample is in English and Persian, including emojis, hashtags and geotag. Please note that personal information has been masked for the purpose of publishing and observing ethical research practices. See Section III-A for additional detail on data anonymisation.

We divided our statistical population (86 profiles) into private and public groups. The eligible participants in the private group should be 25 years of age and older, have met with the lead psychologist for at least 15 hours of counselling sessions, have an Instagram account, and have consented to provide their contact details and profile data. A one-page informative summary (see Appendix I) was given to potential participants to inform them of the purpose of the research and how the data would be used. Finally, ten potential participants

(5 male, 5 female) who were fully informed and returned informed consent were included in the private group. Clinical assessment and in-person diagnosis were, therefore, available in the assessment of their Instagram profiles.

For the public group, we targeted users in Iran and Spain, who only use Instagram for personal purposes. Taking the recommendation of the lead psychologist and our previous experience in data collection [30, 31, 32], we ignored profiles belonging to celebrities, influencers and/or business sectors. Subjects are Iranian and Spanish, between 25 and 50 years of age, with a gender ratio (male/female) of 1.7, *i.e.*, 49 males and 27 females. The public group includes individuals who have never met our leading psychologist. The medical diagnostic codes in [13] have, therefore, been used for annotation.

Two expert psychologists (Persian native speakers) who have been trained in Reality Therapy [13] have reviewed all the visual content of each profile in the dataset in both phases to provide a consensus annotation per profile. The primary reason for excluding textual content from the ground-truthing process is that nuances in language are mainly understandable by native language speakers. Nonetheless, visual content often evokes common concepts [11], which helps to minimise implicit bias. For the multi-label aspect of this study, each profile was labelled with a subset expressed as $L = \{Survival, Belonging, Power, Freedom, Fun\}$, except for the empty set. In addition to the labels, expert psychologists outlined their evidences for perceiving the basic needs of each profile. These free-form descriptions, which do not involve technical codes, have been translated into English by an expert translator and a bilingual speaker verified the accuracy of the translations. The distributions of the perceived needs are shown in Figure 2.

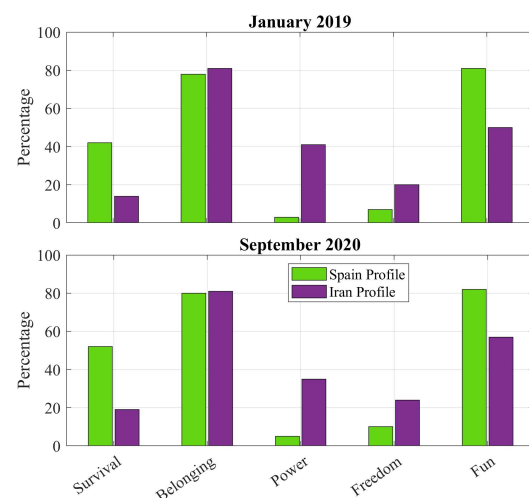


Fig. 2. Diversity of labels per country. The top row shows the percentage of each basic need in January 2019 and the bottom row shows these statistics in September 2020.

In the data acquisition, we used random sampling to ensure that our samples could be representative of the population. Since the sample size (86 Instagram profiles) is far less than the total number of Instagram accounts (over one billion

active users⁵), an unintentional sampling bias could occur. Following the suggestion in [33], we hypothesised that if the distribution of basic needs for Iranian and Spanish users differs significantly, the probability of sampling bias is insignificant.

We run Welch's T-test [34, 35] with the $\alpha = 0.05$ significance level to validate this hypothesis. The Welch T-test for Phase 1 (January 2019) has the degree of freedom $DF = 7.24$ and results in $p = 0.96$, $p(x \leq T) = 0.51$, where the test statistic $T = 0.04$ is in the 95% critical value accepted range of $[-2.34, 2.34]$. For Phase 2 (September 2020) the test has $DF = 7.13$ and results in $p = 0.9$, $p(x \leq T) = 0.55$, where test statistic $T = 0.12$ is in the 95% critical value accepted range of $[-2.35, 2.35]$. Since p -values are significantly greater than α , our hypothesis is acceptable.

A. Ethical procedure

The tools developed to analyse social media data for assessing the psychological dimensions of individuals pose two major ethical concerns, among others: the potential application for which the tool has been developed, and data protection practices and policies.

The architecture proposed in this paper is the initial prototype for exploring whether contextual cues from Instagram data could be used to assess a new dimension of research in the field of affective computing, i.e., the automatic prediction of the five basic needs based on Glasser's choice theory. The algorithm presented in this paper is intended to be used solely for private and non-profit purposes. Although a similar approach can be used in prospective clinical applications (e.g. identifying the source of personal unhappiness to prevent anxiety or depression), this will require further technical and clinical contribution to this particular dimension and would entail written consent and authorisation from patients who would like to participate in this type of study.

The data management protocol followed was as restrictive as possible. A one-page informative summary of the project had been given to the users, while including only those who returned the signed and written informed consent⁶ in the private group. Participants were given the option to contact the corresponding author via e-mail in order to address/resolve any questions or concerns.

For the public group, we used only the data of the users owning public Instagram profiles. To ensure that the user did not change their mind, we downloaded the data while keeping it for just one month. Then we asked two expert psychologists to review the visual content of the profiles, in which there was no link to personal information, to provide a consensus label for each profile. We trained the proposed pipeline using these labels and wiped out all data from our secure server once the training was completed. We only kept the links to these profiles. Therefore, if users delete data or change any settings

in the future, our access to this content through links will be affected in compliance with the privacy policy established by each user.

The project was submitted to the Open University of Catalunya Ethical Committee (IRB). Having considered the ethical implications concerning human experimentation and the processing of personal data and the procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects, the committee approves to report aggregated results from the experiments (i.e., the average score) in order to ensure the integrity and dignity of the participants and avoid the possibility of identifying the original users or any information from their profiles.

IV. METHODOLOGY

Users can use both visual and textual content in the Instagram profile to create a personal photo album. While the owner of the profile can exclusively use the visual content, the textual content can be used by the owner and followers to interact. Photos do not have language-related restrictions and can evoke common concepts [11]. For this reason, the evaluation of Instagram photo albums by psychologists to perceive the basic needs of the user is less challenging than the evaluation of textual data. An expert can see from an image, for example, whether the people appearing in that image are having fun. Yet, users may use textual content to strengthen the representation of their feelings in their profile. Thanks to significant advances in natural language processing, in this research, we can address linguistic challenges and use both modalities in the assessment of the Instagram profile.

We proposed a multimodal and multi-label classification for perceiving of the five basic needs in accordance with Glasser's choice theory. For visual content representation, we use the Places-CNN [36] scene descriptor and the YOLO-base object detector [21], which have been trained with the Places2 [22] and the Microsoft COCO datasets [23], respectively. To represent textual content, we fine-tune the FastText embedding model [24] with our dataset, which includes mainly English, Spanish, Persian, and Turkish words. In the proposed architecture, we contribute to *Bag-of-Content* (BoC) module to fuse and minimise the dimensionality of the multimodal features in the latent space. The details of the proposed methods are described in the following. Figure 3 shows the proposed architecture.

A. Visual content representation

In order to provide a comprehensive semantic understanding of the image, we first identified the scenes in each image. To estimate the probabilities of 365 categories of places, such as 'lake natural', 'restaurant', 'downtown' and 'train station platform' we used Places-CNN [36]. Places-CNN used the GoogLeNet backbone and was trained with data from Places2 [22]. For each image, categories with a probability of 0.01 or higher were chosen as possible scenes. This process was repeated for all images of a profile, and all possible scenes were concatenated to represent a profile with a list of

⁵<https://www.statista.com/topics/1882/instagram/>

⁶The consent form was based on the information included in the Horizon 2020 Manual/Ethics "GUIDANCE FOR APPLICANTS-INFORMED CONSENT" published by the European Commission, Research Directorate-General, Directorate L — Science, Economy, and Society, Unit L3 — Governance and Ethics. The consent forms were also translated into Persian and given in the written form so that they can be signed.

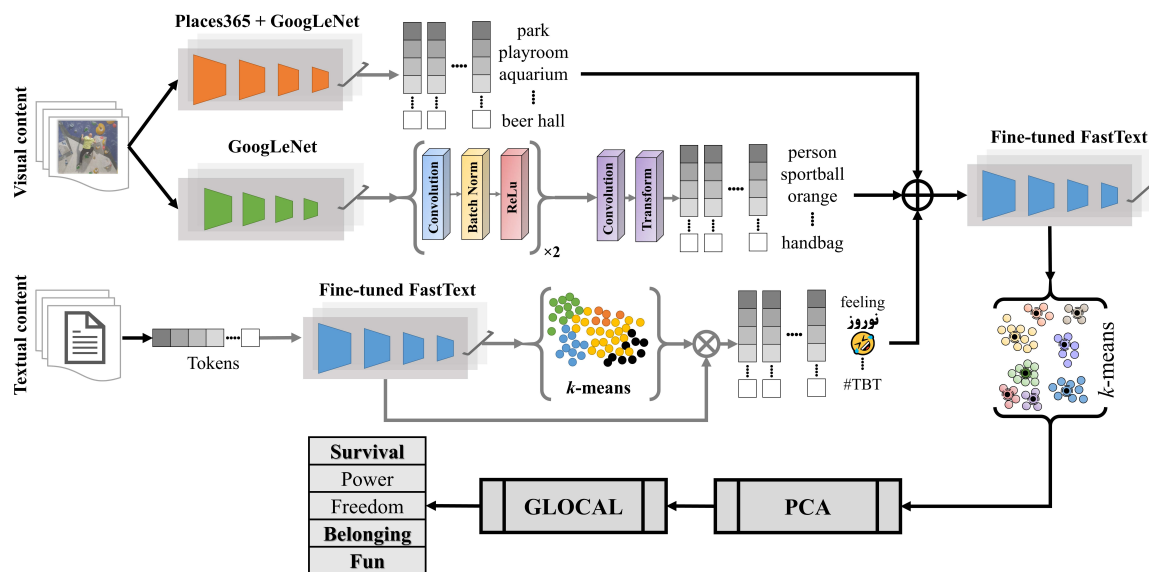


Fig. 3. The pipeline of the proposed approach to predict the basic needs of Instagram users. We used Places-CNN [22] and the modified YOLO-based object detector [21], which were trained with the Places2 [22] and the Microsoft COCO datasets [23], respectively, to extract places and objects. We fine-tuned the FastText embedding model [24] for all words in our data to map each word into a numerical representation. Outputs of these high-level descriptors are integrated by using the proposed BoC to build a bimodal semantic dictionary for each profile. This dictionary is then fed into the GLOCAL multi-label classifier [25] to predict the five basic needs. Please note that in this illustration, we used \oplus to show the appending of all words together, \otimes to show the numerical vector mapping to words, and add \bullet inside each cluster centre.

scenes (S_1). Note that for each profile, the S_1 size is different and may include duplicate place tags. Two examples of scene descriptions are given in Figure 4.

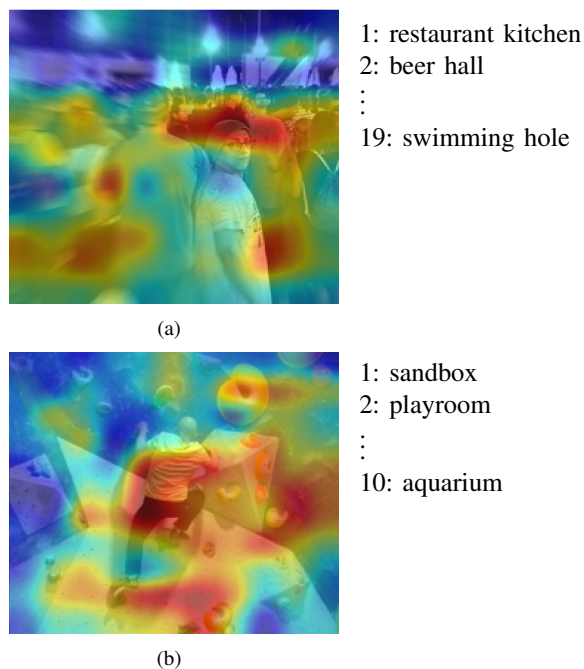
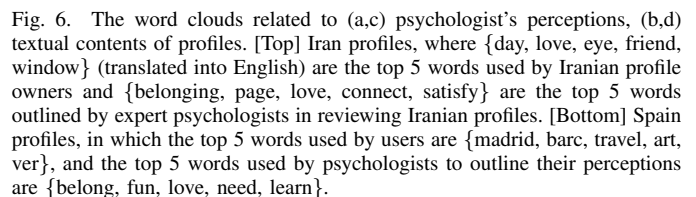
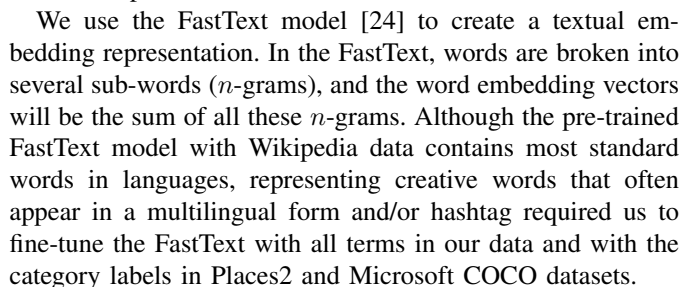


Fig. 4. Places-CNN with GoogLeNet backbone [36] trained with Places2 dataset [22] is used to classify scenes in images. Only categories with a probability of 0.01 or higher were considered. (a) An outdoor and (b) indoor activity with potential tags like 'restaurant kitchen', 'beer hall' and 'playroom'. To highlight the areas considered by CNN, the results were plotted with transparency over the original image.

We used you-only-look-once (YOLO)-based architecture [21] to detect and extract objects. YOLO used the entire top-most feature map to predict confidences for multiple categories of objects at a single stage. The basic idea of YOLO is to divide the input image into an $(p \times p)$ grid. If the centre of an object falls into a grid cell, the grid cell is responsible for detecting the object. YOLO object detection consists of a feature extraction network, followed by a detection sub-network. Here, we used the GoogLeNet backbone in feature extraction network and modified the detection sub-network to address requirements of our study.

In the detection sub-network, we created two groups of serially connected convolution, ReLU and batch normalisation layers. In the first convolution layer, we set the filter size to (7×7) to match the number of channels in the output of the feature extraction layer. The second convolution layer is twice the size of the first layer to allow the model to detect small objects better. These layers are followed by a transform layer with 7 anchor boxes. Anchor boxes extract activations of the last convolutional layer and align predicted bounding boxes with the ground-truth. We trained detection sub-network with Microsoft COCO dataset [23]. This dataset contains 300,000 properly segmented images with an average of 7 object instances out of a total of 80 categories per image. Figure 5 shows the outputs of the proposed object detector. The label of all the objects detected in each image has been saved in a list. This process has been repeated for all images in a profile, and all lists have been concatenated to represent the profile in the S_2 object list.



For each feed, we apply the fine-tuned FastText to each word to transform it into a numerical vector with an embedding dimension of 300. Suppose that each feed contains M words, the representative matrix for this feed has the size of $M \times 300$. We then apply k -means [40] with $k = 5$ (analogous to 5 basic needs) to rank this matrix. To select the most informative rows, the items belonging to the densest cluster are stored in a list. This process is repeated for all feeds in a profile to represent the textual content of the profile with the concatenation of all lists as S_3 . S_3 has a numerical representation while S_1 and S_2 have categorical representations. To harmonise these three representations, we use *vec2word* tool [41] to convert the numerical vectors of S_3 into the categorical one.

C. Bag of Content: A semantic map from visual to textual domain

We borrowed the idea of *Bag-of-Content* (BoC) from natural language processing, where bag-of-words is used to simplify the representation of a document by means of independent words histogram [42]. In this research, we used high-level descriptors to map both the textual and visual content of the profile into three sets of terms, *i.e.*, S_1, S_2, S_3 . Since each list has a different cardinality, a uniform codebook can not be constructed by concatenating these three lists. In addition, the calculation of terms' frequency does not necessarily guarantee the best representation since the relevance and importance of terms are not recorded (for instance, see Figures 4 and 5).

We proposed *Bag-of-Content* to build a codebook that represents both the visual and the textual modalities of a profile for the training of a multi-label classifier. [Algorithm 1](#) presents the main steps to build a BoC that can be easily extended

to include additional modalities. In the proposed BoC, we first concatenate the high-level descriptors S_i , $i = \{1, 2, 3\}$ and then use the fine-tuned FastText word embedding model to transform the list of tokenised documents to sequences of numerical vectors. The vectors are then clustered into λ clusters using the k -means algorithm, with the cluster centres serving as the representative dictionary. We used Principal Component Analysis (PCA) [43] to map the $\lambda \times 300$ representative dictionary into a d -dimensional feature vector, where $d \ll \lambda$ is the number of dimensions. This low-dimensional representation can handle the small sample size, the data set's imbalanced characteristics, and the possibility of the curse of dimensionality when training the classifier.

Algorithm 1: Bag of Content.

Input : S_i – List of terms, containing the output of the i -th high-level descriptor ($i = 1, 2, 3$),
 λ – Number of clusters for BoC,
 d – BoC dimension.

Output: BoC – Representative codebook of profile.

```

1  $\mathcal{P} \leftarrow \bigcup_{i=1}^3 S_i$ ;
2  $n = |\mathcal{P}|$ ;
  //  $n$ : Number of elements in  $\mathcal{P}$ .
3  $\mathcal{V}_{n \times 300} \leftarrow \text{FastText}(\mathcal{P})$ ;
  //  $\mathcal{V}$ : Numerical representation of  $\mathcal{P}$ 
4 Apply  $k$ -means ( $\mathcal{V}_{n \times 300}$ ,  $\lambda$ ) and add cluster centres
  to  $\text{codebook}_{\lambda \times 300}$ ;
5 if  $d > \min(n, 300)$  then
6    $d = \min(n, 300)$ ;
7 end
8  $\text{BoC}_{\lambda \times d} \leftarrow \text{PCA}(\text{codebook}_{\lambda \times 300}, d)$ 
9 return  $\text{BoC}_{\lambda \times d}$ 

```

D. Multi-label classification

Labelling Instagram profiles based on choice theory is naturally linked to more than one class label. Labels that can imply overlap or conflict with the basic needs of the user and others. We use the Multi-Label Learning with GLObal and loCAL Label Correlation (GLOCAL) approach [25] in this research to take advantage of both global and local label correlations.

Let the set of l class labels be expressed in the form of $C = \{c_1, \dots, c_l\}$. The d -dimensional feature vector of an instance is denoted by $x \in X \subseteq \mathbb{R}^d$, and the ground-truth label vector is denoted by $y \in Y \subseteq \{-1, 1\}^l$ where $[y]_j = 1$ if x has the class label c_j and -1 otherwise. GLOCAL exploits the regularisation of global and local manifolds as well as low-rank decomposition to utilise both global and local label correlations. The label matrix in GLOCAL is decomposed to two low-rank Laplacian matrices to substitute missing-label instances with the label correlation to minimise the reconstruction error in the output of the classifier (\hat{Y}).

V. RESULTS

We evaluate our approach in three experiments: (1) a comparison with the subjective test; (2) an analysis of the impacts of λ and d (PCA dimension) on BoC and classifier; (3) an ablation study to demonstrate the contribution of each modality to the proposed architecture. The following sections include explanations of these experiments.

A. Subjective test

We perform a subjective test due to the unavailability of such a systematic study to be compared with our approach. We also analyse the alignment of non-experts' opinions with the experts' to check the feasibility of including workers from crowd-sourcing sites in the annotation task.

In this test, eight bilingual volunteers were asked to annotate the visual content of two Instagram profiles, each with an average of 286 feeds, using the choice theory. We divided the participants into two groups with a gender ratio of 1. Four Persian/English speakers and four Spanish/English speakers were assigned to $G_{Persian}$ and $G_{Spanish}$, respectively. We randomly selected four public profiles belonging to Iranian users and four public profiles belonging to Spanish users from our data set and added them to TG_{Iran} and TG_{Spain} , respectively. We have done this to avoid unintentional bias to possible personal bonding. We also asked all of the volunteers to review the Instagram profile of the main author. Since all volunteers have known the main author personally, the review of this profile was somewhat similar to the assessment of a profile by expert psychologists when an in-person diagnosis is available. The results of the subjective test are shown in Table I.

TABLE I
COMPARISON OF NON-EXPERTS' AND EXPERTS' OPINIONS IN PERCEIVING BASIC NEEDS OF NINE PROFILES. THE MEAN INTRACLAS CORRELATION COEFFICIENT AND CONFIDENCE LEVEL ARE $\bar{r} = 0.22$ AND $\bar{p} = 0.64$, RESPECTIVELY.

	Non-experts					Experts					ICC(A,1)	Confidence level (p)
	Fun	Belonging	Power	Freedom	Survival	Fun	Belonging	Power	Freedom	Survival		
Profile 1			×	×		×					-0.50	0.21
Profile 2	×		×			×				×	0.60	0.92
Profile 3		×				×	×				0.60	0.92
Profile 4	×			×		×					0.60	0.92
Profile 5	×	×	×	×		×	×	×	×		1.00	1.00
Profile 6		×	×			×	×				0.20	0.62
Profile 7	×			×				×	×		0.20	0.62
Profile 8		×			×	×	×				-0.50	0.21
Mutual profile	8/8	7/8	4/8	4/8	2/8	×	×			×	-0.17	0.38

Since the participants were unaware of the choice theory, we provided them with a summary of the five basic needs and a concise Infographic. This infographic (see Appendix I) was originally published on the website of the Glasser Institute for Choice Theory. We did not compel volunteers to follow a particular protocol to review profiles. In fact, it was their own choice to first look at the Instagram user's bio, and then review feeds or to follow a different order.

The disparity of opinions is evident in Table I. **However, we assess the consistency of the non-expert and expert observers' quantitative measurements by measuring the intra-class correlation coefficient (ICC) [44]. We chose the 'A-1' type to measure ICC because we were interested in measuring the absolute agreement between the two raters in the presence of random residual errors and the two raters' systemic errors. The mean intra-class correlation coefficient and confidence levels are $\tilde{r} = 0.22$ and $\tilde{p} = 0.64$, respectively, implying a low intra-class correlation.** These statistics also indicate that, contrary to the suggestion made by [45, 46], the annotation of data on this specific topic, which requires psychological expertise, cannot benefit from the recruitment of workers from crowd-sourcing sites.

B. Architecture details

Experiments were performed on the NVIDIA RTX 2080 GPU with 8 GB of memory in MATLAB 2020a using Deep Learning and Text Analytic toolboxes. In the training of GoogLeNet backbone [36], stochastic gradient descent with momentum algorithm [47] was used to update learning parameters with initial learning rate and momentum of 0.001 and 0.9, respectively. We fed the network with a batch size of 8 and the optimisation stopped after 20 epochs. Both the scene descriptor and the YOLO-based object detector used these hyperparameters.

The YOLO-based object detector consists of feature extraction and object detection modules. In the feature extraction module, we used GoogLeNet backbone which was trained with Microsoft COCO data [23]. All input images were scaled to 224×224 . We used the last two layers with output sizes of 14×14 and 7×7 , respectively, in the object detection module. We excluded pooling layers to preserve the spatial features and replaced them with 16×16 and 32×32 strides, respectively. The size of the stride is the size of the input image to the size of the output feature map.

We used anchor boxes to detect all objects included in the cells of the 7×7 and 14×14 grid size feature maps. The number of anchor boxes depends on the dataset characteristics. In [21], the authors proposed to use the k -means clustering algorithm to estimate the number of anchor boxes. They replaced the direct Euler distance metric by intersection over union (IOU) in k -means to select a bounding box with the highest detection probability. We used the *estimateAnchorBoxes* tool [21] to estimate the number and properties of the anchor boxes, which led to the selection of 7 anchor boxes with an IOU threshold of 0.4. In order to retain the best bounding box, we used Non-Maximal Suppression (NMS) and set the NMS threshold to 0.6. Thus, all predicted bounding boxes with a detection probability of less than 0.6 have been omitted.

There are 80 categories of objects in the Microsoft COCO dataset. As a consequence, each anchor box (x, y, w, h, s, c) has 85 properties where the bounding box properties are represented by (x, y, w, h) , s is the detection score, and c refers to the number of classes. Anchor boxes only added to the final layer of architecture in which each cell includes $7 \times 85 = 595$ elements, making $8 \times (14 \times 14) \times 595$ predictions with a batch size of 8.

We built a dictionary using all the words in our dataset to fine-tune the FastText model. We kept words that appeared at least three times in the training set. We used a context window whose size was uniformly sampled from 1 to 5, and preserved the default values of *skipgram* = 0.025 and *cbow* = 0.05 in *word2vec* [41]. We also set the rejection threshold to 10^{-4} to sub-sample the most frequent words (see [41] for more details). We downloaded Wikipedia dumps of Spanish, Persian, English, and Turkish languages. We normalised the raw Wikipedia data using a semantic vector tool⁷ before merging it to our dataset. The training was carried out with a five-pass over a dataset randomly shuffled in each pass. The training and test sets in each pass contain 80% and 20% of the data, respectively.

C. Experiments

We used three policies to train and evaluate the GLOCAL classifier in our experiments. In total, 38,356 images were sent to image descriptors, and 6,943,050 words were used to refine the FastText model. In the first policy, We used the cross-validation strategy (leave one subject out) to handle the small sample size. To minimise statistical uncertainty, the results were averaged across independent repetitions. In the second policy, we trained GLOCAL with 77 samples (35,382 images and 6,619,614 words) and validated it with the nine samples used for the subjective test. This 90/10 split helps us to compare the proposed approach to the subjective test. In the third policy, we trained GLOCAL with all of the samples collected in Phase 1 (30,824 images and 5,594,880 words) and tested it with samples (7,532 images and 1,348,170 words) added to the dataset in Phase 2.

1) Evaluation metrics

At the suggestion of Pereira et al. [48], we used Hamming Loss, Coverage, Example-Based Accuracy, Ranking Loss, and F-Measure to report the performance of the GLOCAL multi-label classifier. These metrics could prevent the presentation of redundant information in the assessment.

- Hamming Loss (HL) is a normalised metric in which a prediction error (when an incorrect label is predicted) and a missing error (when a relevant label is not predicted) are considered for all classes. It can be calculated by $HL(H, X) = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \Delta \hat{Y}_i|}{|C|}$, where H is the generated model by the multi-label classifier, N is the number of test data, and Δ is the symmetrical difference between the two sets, similar to the XOR operation in Boolean logic.
- Coverage (Cvg) counts on average the steps to be taken in the ranked list of labels to cover all the relevant labels

⁷<https://github.com/PrincetonML/SemanticVector>

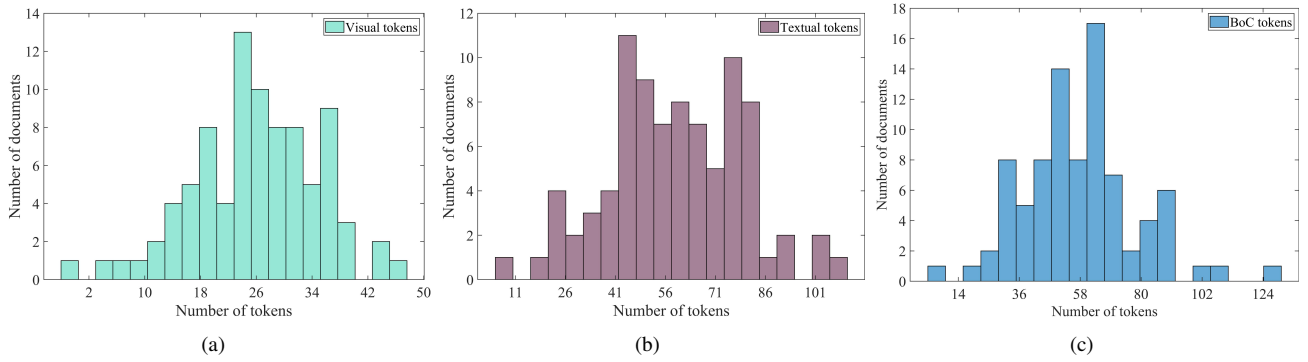


Fig. 7. Terms count histogram for (a) visual-extracted terms, (b), textual-extracted terms, and (c) BoC-extracted terms.

of the example. It can be calculated by $\text{Cvg}(H, X) = \frac{1}{N} \sum_{i=1}^N \max(r_i(c)) - 1$, where $r_i(c)$ is the rank position of the label c . The most relevant label has the highest rank and the least relevant label has the lowest rank (l).

- Example-based Accuracy (EbA) expresses the overall effectiveness of a classifier, given by $\text{EbA}(H, X) = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|}$.
- Ranking Loss (Rkl) calculates the frequency of irrelevant labels that are ranked higher than relevant labels, given by $\text{Rkl}(H, X) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i| |\hat{Y}_i|} |\{(c_a, c_b) : r_i(c_a) > r_i(c_b), (c_a, c_b) \in Y_i \times \hat{Y}_i\}|$.
- F-Measure is the harmonic mean of Precision and Recall, which is calculated by $\text{F1}(H, X) = \frac{1}{N} \sum_{i=1}^N \frac{2|Y_i \cap \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|}$, where we report it in percentage F1(%).

To provide more insight into the performance of predictive model with respect to different dimensions of the basic needs, we calculated precision, recall, and F1, in the third policy using Eq. 1.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, & \text{Recall} &= \frac{TP}{TP + FN}, \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \quad (1)$$

where TP , FP and FN stand for true positive, false positive and false negative, respectively.

2) Impact of cluster size and PCA

The number of clusters (λ) and PCA dimensions (d) contribute to the performance of the proposed method. We first plotted the number of terms in the training data (see Figure 7) and found that the distribution of BoC terms is within $[30, 70]$. Then to find the most appropriate cluster number (λ) and to analyse the impact of applying PCA to the representative dictionary, we trained GLOCAL with the first policy as a function of (λ, d) and plotted the Hamming Loss.

Figure 8 shows that the minimum values of hamming loss (with a minor difference) obtained for the two combinations of $(\lambda, d) = \{30, 2\}$ and $(\lambda, d) = \{70, 1\}$. We measured the variance of the principal component in order to select the appropriate d , where $\lambda = \{30, 70\}$. We found that $d = 2$ is capable of explaining $\{89.6\%, 49.6\%\}$ of the total variances in relation to the corresponding λ (see Figure 9).

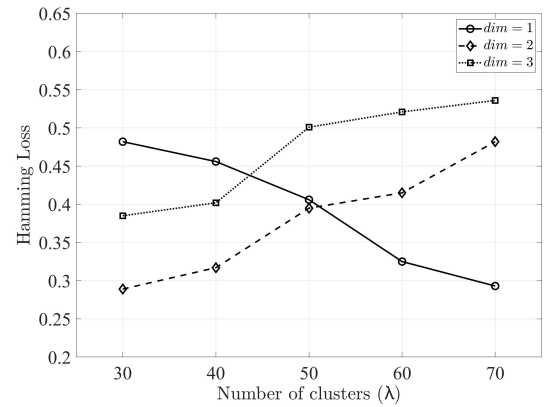


Fig. 8. Hamming loss of GLOCAL as a function of $\{\lambda, d\}$. GLOCAL achieved $\text{HL} = 0.293$ with $\{\lambda, d\} = \{70, 1\}$ and $\text{HL} = 0.289$ with $\{\lambda, d\} = \{30, 2\}$. In $\{\lambda, d\} = \{30, 2\}$ vectors were concatenated before passing to the classifier.

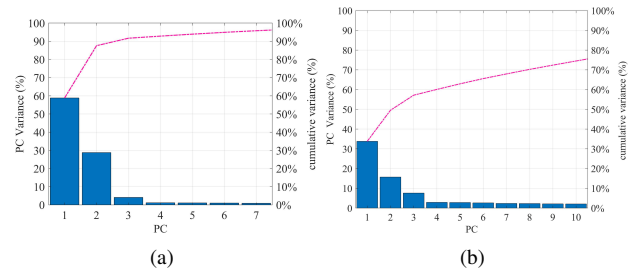


Fig. 9. Pareto plot for variance percentages of the PCA dimensions, where (a) $\lambda = 30$ and (b) $\lambda = 70$. In this plot, PC stands for the principal component.

Table II shows the performance of the proposed method in the first training policy with $(k, d) = \{(70, 1), (30, 2)\}$.

TABLE II
PERFORMANCE METRICS OF GLOCAL TRAINED BASED ON THE FIRST POLICY WITH $(\lambda, d) = \{(70, 1), (30, 2)\}$. WE SHOW THE MEAN OF METRICS AT 95% CONFIDENCE INTERVALS.

Metric	(70, 1)	(30, 2)
Hamming Loss	0.32 ± 0.06	0.28 ± 0.07
Coverage	3.23 ± 0.06	3.79 ± 0.05
Ranking Loss	0.44 ± 0.09	0.35 ± 0.01
Example-Based Acc.	0.67 ± 0.08	0.69 ± 0.11
F-Measure (%)	68.94 ± 1.20	76.34 ± 1.80

The second training policy attempts to compare the BoC-

based multi-label classification with each of the visual and textual modules. We followed the split of 90%-10% where the test set comprised of profiles that were used in the subjective test. The results presented in Table III indicate that high-level textual descriptors help to better understand the five basic needs than visual descriptors. The two key explanations for understanding this effect are: 1) The user can express his/her needs and choices in the feed or comments and post an image without a semantic association to the text content. 2) The user has the ability to interact with the followers through text. The possibility of self-disclosure of needs is higher in this two-way conversation. The words cloud in Figure 6 implicitly confirm these observations. However, we have shown in the Ablation study (see Section V-D) that the impact of visual content is undeniable, where the absence of visual modality leads to the attenuation of Bag-of-Content.

TABLE III
PERFORMANCE OF GLOCAL ON PREDICTING BASIC NEEDS FOR $(k, d) = (30, 2)$, WHERE THE PROPOSED PIPELINE IS TRAINED WITH THE SECOND POLICY. WE COMPARE THE PROPOSED METHOD WITH EACH OF THE VISUAL AND TEXTUAL DESCRIPTORS.

Method	HL	Cvg	Rkl	EbA	F1 (%)
Bag of Content	0.06	2.33	0.09	0.66	84.07
Places-CNN	0.10	2.55	0.21	0.44	80.26
YOLO	0.28	3.33	0.35	0.22	74.49
FastText	0.08	2.44	0.15	0.55	82.48

In the third policy, we first examined the bias effect and generalisability of the proposed approach for predicting labels from new profiles. We also evaluated the dependency of the proposed approach on the size of the data. To analyse the bias, we train the proposed architecture with data collected in Phase 1 and test it with data collected in Phase 2 (see Table IV and V). To analyse the scale dependence, we used the cross-validation strategy (leave one subject out) only for the portion of the data collected in Phase 2. The results are shown in Table VI.

TABLE V
PERFORMANCE OF THE PROPOSED APPROACH CONSIDERING CLASS-WISE RECALL, PRECISION, AND F1 SCORE. HERE, TP, TN, FP AND FN STAND FOR TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE AND FALSE NEGATIVE, RESPECTIVELY.

	TP	TN	FP	FN	Precision	Recall	F1-score
Survival	29	29	10	18	0.74	0.62	0.67
Belonging	17	55	9	5	0.65	0.77	0.71
Power	12	45	15	14	0.44	0.46	0.45
Freedom	10	48	18	10	0.36	0.5	0.42
Fun	62	10	7	7	0.90	0.90	0.90

TABLE IV
EXAMINING THE BIAS EFFECT AND GENERALISABILITY OF THE PROPOSED APPROACH IN THE THIRD POLICY. WE COMPARE THE PROPOSED APPROACH WITH THE VISUAL AND TEXTUAL DESCRIPTORS.

Method	Hamming Loss	Coverage	Ranking Loss	Example-Based Accuracy	F-Measure (%)
Bag of Content	0.04	2.33	0.06	0.77	94.07
Places-CNN	0.08	2.55	0.15	0.55	90.26
YOLO	0.22	3.33	0.35	0.33	74.49
FastText	0.06	2.44	0.09	0.66	92.48

D. Ablation study

In this section, we examine the performance and role of each module in the task of multi-label classification (see Table 6). We trained the classifier with five passes over the training set containing 77 samples (approximately 90% of dataset size), which is randomly shuffled in each pass. Details of this experiment are as follows:

- 1) We removed the proposed Bag-of-Content. To represent each profile, we applied steps 1-3 in Algorithm 1 to \mathcal{P} and calculated the average of all vectors in \mathcal{V} .
- 2) We disabled the YOLO-based object detector to evaluate the contribution of the scenes. However, we used BoC in the architecture to integrate feature spaces.
- 3) We used the Places-CNN scene descriptor and the YOLO-based object detector separately to represent each profile image. This process was repeated for all profile images, and two frequency histograms of the predicted tags were produced. To deal with mutual exclusion, these histograms are then weighted by the tags score. For our dataset, $\Upsilon_{86 \times 365}^{Places}$ and $\Upsilon_{86 \times 80}^{YOLO}$ represent the visual feature spaces.
- 4) To evaluate the contribution of the FastText embedding model, we created a histogram of words occurrence for each Instagram profile. For our dataset, $\Upsilon_{86 \times 300}^{FastText}$ represents the textual feature space.
- 5) To assess the impact of k -means on the integration of these modalities into the BoC, we removed this module and applied PCA to the linear concatenation of \mathcal{P} entries. As a result, all profiles were represented by the $\Upsilon_{86 \times 86}^{Fusion}$ feature space.

The comparison of results given in Tables III — VII indicates that both visual and textual modalities, particularly when used together, contribute to the perception of basic needs. However, the elimination of BoC reduces the efficiency of the proposed method for two reasons: (1) overfitting due to the high ratio of the feature dimensions to the number of samples; and (2) attenuating the multimodal data semantic relationship.

Although Table VII indicates that the YOLO-based object detector does not make a significant contribution to the perception of basic needs, its elimination reduces performance metrics due to the attenuation of the semantic relationship. To elaborate, suppose a user posts on Instagram a photo of a birthday party, and another user shares a picture of a camp with friends. The semantic relationship is what allows the proposed pipeline to differentiate between the need for *Belonging* and *Fun* in both cases where 'person' is the dominant object.

TABLE VI

ANALYSING THE SCALE DEPENDENCY OF THE PROPOSED APPROACH IN THE THIRD POLICY. WE COMPARE THE PROPOSED APPROACH WITH THE VISUAL AND TEXTUAL DESCRIPTORS.

Method	Hamming Loss	Coverage	Ranking Loss	Example-Based Accuracy	F-Measure (%)
Bag of Content	0.22 ± 0.01	3.77 ± 0.31	0.38 ± 0.02	0.22 ± 0.02	70.74 ± 0.55
Places-CNN	0.26 ± 0.09	3.66 ± 0.56	0.50 ± 0.07	0.33 ± 0.03	61.66 ± 0.21
YOLO	0.35 ± 0.07	4.11 ± 0.77	0.62 ± 0.10	0.11 ± 0.02	57.28 ± 0.61
FastText	0.31 ± 0.02	3.77 ± 0.29	0.51 ± 0.04	0.22 ± 0.03	59.31 ± 0.84

TABLE VII

PERFORMANCE OF GLOCAL TO PREDICT BASIC NEEDS IN THE ABLATION STUDY. \odot IS USED TO SHOW THE ELIMINATION OF A MODULE. WE SHOW THE MEAN OF METRICS AT 95% CONFIDENCE INTERVALS.

Method	Dimension	Hamming Loss	Coverage	Ranking Loss	Example-Based Accuracy	F-Measure (%)
Proposed	30	0.28 ± 0.03	3.33 ± 0.23	0.35 ± 0.05	0.68 ± 0.02	71.34 ± 1.64
Proposed \odot YOLO	30	0.32 ± 0.08	3.77 ± 0.11	0.37 ± 0.10	0.59 ± 0.08	68.94 ± 2.10
Proposed \odot BoC	300	0.46 ± 0.06	4.87 ± 0.53	0.40 ± 0.05	0.55 ± 0.09	61.66 ± 0.35
$\Upsilon^{FastText}$	300	0.51 ± 0.03	5.25 ± 0.25	0.58 ± 0.08	0.46 ± 0.05	52.33 ± 0.83
Υ^{Places}	365	0.56 ± 0.06	5.80 ± 0.18	0.62 ± 0.05	0.38 ± 0.06	50.46 ± 1.25
Υ^{YOLO}	80	0.67 ± 0.09	6.32 ± 0.24	0.66 ± 0.18	0.22 ± 0.12	44.38 ± 2.47
Υ^{Fusion}	86	0.49 ± 0.03	5.25 ± 0.25	0.44 ± 0.08	0.50 ± 0.05	58.71 ± 0.75

VI. CONCLUSION

A relevant part of our time is consumed by sharing multimodal data on social media platforms such as Instagram, Facebook and Twitter. In social media, the specific way users express themselves can provide important insights into their behaviours, personalities, perspectives, motivations and needs. The primary concern is how truly representative social media is, given the results of studies that indicate a strong correlation between the shared content of users on social networks and their mental states [4, 5, 6, 7]. Are users revealing their true needs and mental states, or they are trying to build a perfect image through their social media profiles to escape from real life?

This concern has always existed in the field of affective computing, where for example, we trust the available tagged data in the training of an emotional recognition model, apart from the fact that the image may represent true or imposed happiness. Glasser's choice theory [13] was built on the basis of our *choices* when we chose to show or pretend happiness. That is why this study focused on choice theory to alleviate bias from this viewpoint and open up another dimension to affective computing. The choice of content to share interests and feelings on Instagram is analogous to the creation of the Personal Picture Album in Glasser's Choice Theory. Such contents can, in any sense, disclose the unmet basic needs of the Instagram user to a psychologist. Identifying unmet needs is the first step for a psychologist who follows reality therapy to help people find solutions for their psychological problems or improve their quality of life.

In this paper, for the first time, we studied how individuals intrinsically contribute to the basic needs by choosing content to share on the Instagram profile. In order to perceive the five basic needs from Instagram accounts, we introduced a multimodal classification system that benefits from state-of-the-art CNN-based visual and textual content descriptors. To capture the conceptual relationship between visual and textual modalities, we also proposed the *Bag-of-Content* (BoC). In BoC, identified scene by Places-CNN [22] and detected objects

by the YOLO-based object detector [21] were integrated with words represented by FastText embedding model [24]. Comprehensive evaluations demonstrated higher performance for the proposed multimodal and multi-label approach compared to the results of subjective test.

It must be noted that the developed classification architecture is intended for private use, not for use by other parties, with the patient's consent to assist psychologists in the identification of early signs of unhappiness. However, our methodology can be adapted to address similar multi-class or multi-label psychological research where multimodal social media data is targeted. Future studies need to address ethical concerns in order to incorporate more data from a wide variety of social media platforms. Also, more attention to cultural adaptations allows key stakeholders to benefit from the results of these studies.

ACKNOWLEDGEMENT

We would like to thank anonymous reviewers for their critical reading and for providing insightful feedback that helped us to improve and clarify this manuscript.

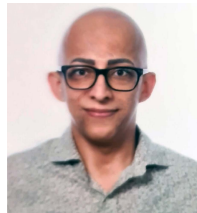
This research was supported by TIN2015-66951-C2-2-R, RTI2018-095232-B-C22 grant from the Spanish Ministry of Science, Innovation and Universities (FEDER funds), and NVIDIA Hardware grant program.

REFERENCES

- [1] D. Scott and B. Happell, "The high prevalence of poor physical health and unhealthy lifestyle behaviours in individuals with severe mental illness," *Issues in mental health nursing*, vol. 32, no. 9, pp. 589–597, 2011.
- [2] M. M. Dehshibi, G. Pons, B. Baiani, and D. Masip, "Vic-som: Visual clues from social media for psychological assessment," *arXiv preprint arXiv:1905.06203*, pp. 1–12, 2019.
- [3] S. S. Khoo and H. Yang, "Social media use improves executive functions in middle-aged and older adults:

- 1 A structural equation modeling analysis,” *Computers in*
- 2 *Human Behavior*, vol. 111, p. 106388, 2020.
- 3 [4] A. G. Reece and C. M. Danforth, “Instagram photos
- 4 reveal predictive markers of depression,” *EPJ Data Sci-*
- 5 *ence*, vol. 6, no. 1, p. 15, 2017.
- 6 [5] D. Muriello, L. Donahue, D. Ben-David, U. Ozertem,
- 7 and R. Shilon, “Under the hood: Suicide prevention tools
- 8 powered by ai,” *Facebook Code*, 2018.
- 9 [6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection
- 10 of depression-related posts in reddit social media forum,”
- 11 *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.
- 12 [7] A. B. Shatte, D. M. Hutchinson, and S. J. Teague,
- 13 “Machine learning in mental health: a scoping review
- 14 of methods and applications,” *Psychological medicine*,
- 15 vol. 49, no. 9, pp. 1426–1448, 2019.
- 16 [8] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and
- 17 J. W. Pennebaker, “The psychology of word use in
- 18 depression forums in english and in spanish: Texting two
- 19 text analytic approaches.” in *ICWSM*, 2008.
- 20 [9] P. Cuijpers, O. Eylem, E. Karyotaki, X. Zhou, and
- 21 M. Sijbrandij, “Psychotherapy for depression and anxiety
- 22 in low-and middle-income countries,” in *Global Mental*
- 23 *Health and Psychotherapy*. Elsevier, 2019, pp. 173–192.
- 24 [10] G. Bernal, J. Bonilla, and C. Bellido, “Ecological validity
- 25 and cultural sensitivity for outcome research: Issues for
- 26 the cultural adaptation and development of psychosocial
- 27 treatments with hispanics,” *Journal of abnormal child*
- 28 *psychology*, vol. 23, no. 1, pp. 67–82, 1995.
- 29 [11] E. A. Ware, S. A. Gelman, and F. Kleinberg, “The
- 30 medium is the message: Pictures and objects evoke
- 31 distinct conceptual relations in parent-child conversa-
- 32 tions,” *Merrill-Palmer quarterly (Wayne State University*
- 33 *Press)*, vol. 59, no. 1, 2013.
- 34 [12] A. T. Beck, *Depression: Clinical, experimental, and*
- 35 *theoretical aspects*. Hoeber Medical Division, Harper
- 36 & Row, 1967.
- 37 [13] W. Glasser, *Choice theory: A new psychology of personal*
- 38 *freedom*. HarperCollins Publishers., 1998.
- 39 [14] B. D. Loyd, “The effects of reality therapy/choice theory
- 40 principles on high school students’ perception of needs
- 41 satisfaction and behavioral change,” *International Jour-*
- 42 *nal of Reality Therapy*, vol. 25, no. 1, 2005.
- 43 [15] P. A. Robey, “Reality therapy and choice theory: An
- 44 interview with robert wubbolding,” *The Family Journal*,
- 45 vol. 19, no. 2, pp. 231–237, 2011.
- 46 [16] O. Massah, F. Farmani, R. Karimi, H. Karami, F. Hoseini,
- 47 and A. Farhoudian, “Group reality therapy in addicts
- 48 rehabilitation process to reduce depression, anxiety and
- 49 stress,” *Iranian Rehabilitation Journal*, vol. 13, no. 1, pp.
- 50 50–44, 2015.
- 51 [17] W. Glasser, *Warning: Psychiatry can be hazardous to*
- 52 *your mental health*. HarperCollins Publishers, 2003.
- 53 [18] A. Al-Kandari, S. R. Melkote, and A. Sharif, “Needs and
- 54 motives of instagram users that predict self-disclosure
- 55 use: A case study of young adults in kuwait,” *Journal*
- 56 *of Creative Communications*, vol. 11, no. 2, pp. 85–101,
- 57 2016.
- 58 [19] H. S. Hwang and J. Cho, “Why instagram? intention
- 59 to continue using instagram among korean college stu-
- 60 dents,” *Social Behavior and Personality: an international*
- journal*, vol. 46, no. 8, pp. 1305–1315, 2018.
- [20] S. Hong, M. R. Jahng, N. Lee, and K. R. Wise, “Do
- you filter who you are?: Excessive self-presentation,
- social cues, and user evaluations of instagram selfies,”
- Computers in Human Behavior*, vol. 104, p. 106159,
- 2020.
- [21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You
- only look once: Unified, real-time object detection,” in
- Proceedings of the IEEE conference on computer vision*
- and pattern recognition*, 2016, pp. 779–788.
- [22] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Tor-
- ralba, “Places: A 10 million image database for scene
- recognition,” *IEEE transactions on pattern analysis and*
- machine intelligence*, vol. 40, no. 6, pp. 1452–1464,
- 2018.
- [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona,
- D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft
- coco: Common objects in context,” in *European confer-*
- ence on computer vision*. Springer, 2014, pp. 740–755.
- [24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov,
- “Enriching word vectors with subword information,”
- Transactions of the Association for Computational Lin-*
- guistics*, vol. 5, pp. 135–146, 2017.
- [25] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, “Multi-label learn-
- ing with global and local label correlation,” *IEEE Trans-*
- actions on Knowledge and Data Engineering*, vol. 30,
- no. 6, pp. 1081–1094, 2018.
- [26] Y. Kim and J. H. Kim, “Using computer vision tech-
- niques on instagram to link users’ personalities and
- genders to the features of their photos: An exploratory
- study,” *Information Processing & Management*, vol. 54,
- no. 6, pp. 1101–1114, 2018.
- [27] K. Kircaburun and M. D. Griffiths, “Instagram addiction
- and the big five of personality: The mediating role of self-
- liking,” *Journal of behavioral addictions*, vol. 7, no. 1,
- pp. 158–170, 2018.
- [28] A. Del Sole, “Introducing microsoft cognitive services,”
- in *Microsoft Computer Vision APIs Distilled*. Springer,
- 2018, pp. 1–4.
- [29] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau,
- F. Yang, M. Pediaditis, and M. Tsiknakis, “Automatic
- assessment of depression based on visual cues: A system-
- atic review,” *IEEE Transactions on Affective Computing*,
- 2017.
- [30] A. Bastanfard, M. A. Nik, and M. M. Dehshibi, “Iranian
- face database with age, pose and expression,” *Machine*
- Vision*, pp. 50–55, 2007.
- [31] M. M. Dehshibi and A. Bastanfard, “A new algorithm for
- age recognition from facial images,” *Signal Processing*,
- vol. 90, no. 8, pp. 2431–2444, 2010.
- [32] M. M. Dehshibi and J. Shanbehzadeh, “Cubic norm and
- kernel-based bi-directional pca: toward age-aware facial
- kinship verification,” *The Visual Computer*, pp. 1–18,
- 2017.
- [33] J. E. Hunter and F. L. Schmidt, *Methods of meta-*
- analysis: Correcting error and bias in research findings*.

- Sage, 2004.
- [34] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.
- [35] H. Farmer, C. Bevan, D. P. Green, M. Rose, K. Cater, and D. Stanton-Fraser, "Did you see what i saw?: Comparing user synchrony when watching 360° video in hmd vs flat screen," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 916–917.
- [36] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.
- [37] S. A. H. Minoofam, M. M. Dehshibi, A. Bastanfard, and P. Eftekhari, "Ad-hoc ma'qeli script generation using block cellular automata," *J. Cell. Autom.*, vol. 7, no. 4, pp. 321–334, 2012.
- [38] M. M. Dehshibi, A. Shirmohammadi, and A. Adamatzky, "On growing persian words with l-systems: visual modeling of neyname," *International Journal of Image and Graphics*, vol. 15, no. 03, p. 1550011, 2015.
- [39] N. Taghipour, H. H. S. Javadi, M. M. Dehshibi, and A. Adamatzky, "On complexity of persian orthography: L-systems approach," *Complex Systems*, vol. 25, no. 2, pp. 127–156, 2016.
- [40] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.
- [41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations (Workshop Poster)*, 2013, pp. 1–12.
- [42] S.-S. Parsa, M. Sourizaei, M. M. Dehshibi, R. E. Shateri, and M. R. Parsaei, "Coarse-grained correspondence-based ancient sasanian coin classification by fusion of local features and sparse representation-based classifier," *Multimedia Tools and Applications*, vol. 76, no. 14, pp. 15 535–15 560, 2017.
- [43] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
- [44] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological methods*, vol. 1, no. 1, p. 30, 1996.
- [45] P. Burnap, W. Colombo, and J. Scourfield, "Machine classification and analysis of suicide-related communication on twitter," in *Proceedings of the 26th ACM conference on hypertext & social media*, 2015, pp. 75–84.
- [46] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 3267–3276.
- [47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [48] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Information Processing & Management*, vol. 54, no. 3, pp. 359–369, 2018.



Mohammad Mahdi Dehshibi is currently a post-doctoral research fellow at Universitat Oberta de Catalunya, Spain. He obtained the PhD from IAU (Iran) in 2017. He was also a visiting researcher at Unconventional Computing Lab, UWE, Bristol, U.K. He has contributed to over 50 papers published in scientific Journals and International Conferences. His research interests include Affective and Unconventional Computing.



Bitia Baiani has 15 years of clinical experience in psychology, and is currently a PhD student in Psychological Therapy at Islamic Azad University (Science and Research Branch), Tehran, Iran. Her research interests include Reality Therapy, Psychology of Personality, Psychoanalysis and Psychosomatic Medicine.



Gerard Pons is a postdoc researcher at Centrum Wiskunde Informatica (The Netherlands). He obtained the PhD from the University of Girona (Spain) in 2014. He was part of the Universitat Oberta de Catalunya (Spain) as a postdoc researcher until 2019. His main research topic is affective computing and emotion recognition using machine learning and computer vision methods.



David Masip is Professor in the Computer Science Multimedia and Telecommunications Department, Universitat Oberta de Catalunya since February 2007 and Director of the Doctoral School since 2015. He is the director of the Scene Understanding and Artificial Intelligence Lab and member of the BCN Perceptual Computing Lab. He studied Computer Vision at the Universitat Autònoma de Barcelona. He received his Ph.D. in 2005 and was awarded for the best thesis in the Computer Science.

Appendix for “A deep multimodal learning approach to perceive basic needs of humans from Instagram profile”

Mohammad Mahdi Dehshibi*, Bitu Baiani†, Gerard Pons‡, David Masip*

*Department of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain

†Department of Psychology, Islamic Azad University, Science and Research Branch, Tehran, Iran

‡Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands

A AN OVERVIEW OF GLASSER’S CHOICE THEORY

According to Glasser’s choice theory [1], basic needs of human are defined as follows (see Figure 1 for a concise Infographic).



Fig. 1. Infographic for explaining 5 basic needs, from Glasser Institute for Choice Theory website [2].

- **Survival:** includes the basic physical and physiological needs such as food, water, air, shelter and clothing. The need to have protection for life is also a part of this need.
- **Belonging:** the need for belonging and an emotional connection to others is fundamental to all human beings. Glasser highlights the essential impact of emotional and sincere relationships on human behaviour. In counselling psychology, belonging is categorised into three categories: (1) social belonging, (2) profession belonging, and (3) family belonging. This classification raises the question: “How does an individual meet the needs of belonging in relation with friends at school, colleagues at work or family and

community members?” Counsellors must foster this context for patients to enable them to develop meaningful and satisfying connections with important people in their lives. In Glasser’s opinion, a lack of emotional engagement, an unhappy relationship, or a relationship with which there is little satisfaction causes significant mental health problems in individuals.

– **Power:** implies the need to gain power, wealth, influence and success, as well as being able to do stuff. This need also includes a sense of accomplishment, development, pride, value, and self-esteem. The need for power is often manifested in competition with those around us. Another part of this need is the desire to perform activities, such as swimming or walking, successfully. Although there is no competition in these activities, accomplishing them can draw a picture of self-confidence in one’s mind. As individuals, our needs for power and belonging can sometimes be at odds with each other. Glasser states that insufficient love is not inherently what ruins a relationship, but what causes a relationship to break up is the struggle for power that manifests itself in the form of control over a marital relationship. He believes that the source of many mental health problems is the inability to obtain a sense of self-worth, which is primarily linked to the early years of life.

– **Freedom:** this need makes it possible for individuals to choose. Moving or migrating from one place to another to express thoughts freely or to have esoteric liberty are manifestations of freedom. Even if environmental deterrents are implemented, by choosing how to respond to the circumstances, people can still preserve their inner freedom. The inability to regulate impulses as well as certain external factors (such as drugs) can restrict this need. What we want as freedom is to live our life voluntarily, to express ourselves freely and to be free from undue external pressures.

– **Fun:** we are perhaps the only creatures that are consciously seeking fun and entertainment. When it comes to travel and leisure, we expend more on happiness and enjoyment than we do on other needs. However, this need is not just about relaxing or enjoying life. In the theory of evolution, fun is a kind of genetic reward that we receive in return for learning.

1 Since we know less than high-level animals when we are
2 born, we need to try harder to learn how to fulfil our needs.

3
4 **B IMPACT OF VISUAL CONTENTS IN PERCEIVING**
5 **BASIC NEEDS**

6 The lead psychologist used the main author’s profile for this
7 paper to exemplify how visual contents help in perceiving
8 basic needs of a user. The main author’s profile is used
9 to avoid ethical issues in light of data protection law. This
10 profile is tagged with $l = \{Survival, Belonging, Fun\}$.



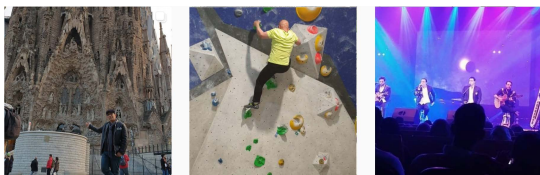
(a)



(b)



(c)



(d)

48 Fig. 2. Sample photos of the corresponding author Instagram profile,
49 including indicators for (a) Survival, (b) profession Belonging, (c) social
50 and family Belonging, and (d) Fun. Please note that for publication
51 purposes and observation of ethical concerns those images that could
52 compromise personal privacy have been masked.

53 In evaluating this profile, the posts can be split into two
54 parts. The first one-third of the posts relate mainly to
55 Survival, while the rest of the feeds, particularly the more
56 recent ones, express the needs of Belonging and Fun. We
57 can see Survival in the feeds that the user shared home-
58 made foods and his cooking (see Figure 2(a)). Slightly earlier
59 feeds include, in particular, selfies where the user involved

in learning-related stuff emphasise a profession-related Be-
longing (see Figure 2(b)). Over time, this need expands
by posting more images with friends and relatives that
refer to social and family-related aspects (see Figure 2(c)).
Pictures showing music festivals, playing sports and travel
(see Figure 2(d)) consider as indicators of Fun.

REFERENCES

[1] W. Glasser, *Choice theory: A new psychology of personal freedom*. HarperCollins Publishers., 1998.
[2] —, “Glasser’s Choice Theory infographic,” <https://wglasser.com/wp-content/uploads/2019/04/Basic-Needs-2.png>, accessed: 2020-04-10.

Summary of Changes

Dear Prof. Elisabeth Andre,

Given the opportunity to revise our paper (*TAFFC-2020-12-0356*) entitled "***A deep multimodal learning approach to perceive humans basic needs from Instagram profiles***," we would like to submit a revised edition of our manuscript as a Research Paper to *IEEE Transactions on Affective Computing*.

We acknowledged the reviewers' and associate editor's comments and responded to them both in the manuscript and in the response letter. To aid in the review process, we presented a "***Response letter***" on the following pages, as well as the highlighted changes in **blue** in the paper.

We hope that the revisions and responses will make the decision to publish this study in *IEEE Transactions on Affective Computing*.

Yours sincerely,
Authors

Evaluation by the Ethics Committee of the UOC

Dr. Marta Aymerich, president of the Ethics Committee of the Universitat Oberta de Catalunya

CERTIFIES

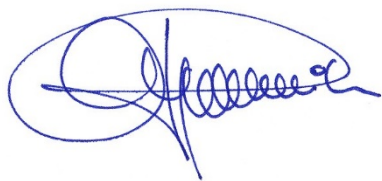
That the Committee has evaluated the project submitted by David Masip Rodó that presents an scientific article to the “IEEE Transactions on Affective Computing Scientific Journal”, and considers that

- The ability of the researchers and their collaborators, and the facilities and resources available are adequate to carry out the scientific article.
- The established experimental protocol ensures the integrity and dignity of the participants.
- The scientific article is adequate to the objectives of the study and the possible risks and discomfort for participants are adequate given the expected benefits.
- The procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects are adequate.
- The researchers of the scientific article will ever respect the obligations derived from the Organic Law 3/2018 on Personal Data Protection and Digital Rights, General Regulation on Data Protection (UE) 2016/679 and the current complementary legislation.

Having met on December 9th 2020, and having considered the ethical implications concerning human experimentation and the processing of personal data, this committee APPROVES the presentation of the aforementioned scientific article.

For the record, I sign this document in Barcelona, December 9th 2020.

Signed:



Dr. Marta Aymerich,

Av. Tibidabo, 39-43
08035 Barcelona – Spain
Tel. +34 93 253 23 00
Fax +34 93 417 64 95

Response to Associate Editor

Comment:

The manuscript has improved post revision.

We would like to thank the reviewers and associate editor for providing insightful comments that definitely helped us to improve the paper's quality. In this letter, we provide detailed answers to all concerns.

Answer: We tried to address all the concerns raised by the reviewers regarding the experimental protocol, and we hope that the paper's quality improved significantly from their feedback.

[AE - Comment#1] The reviewers still have important questions regarding the experimental protocol.

Answer: Regarding the experimental protocol:

Reviewer #1 was concerned about: (1) bias in data annotation due to understanding linguistic nuances that can only be interpreted by the specific native language speaker and (2) the availability of results without PCA in BoC for comparison purposes.

These concerns were addressed thoroughly in **[R#1 - Comment#1]** and **[R#1 - Comment#2]**

Reviewer #2 was concerned about three things: (1) ethical problems, (2) bias in results due to the small sample size, and (3) the efficiency of the proposed approach in the face of an uneven and evolving distribution of needs.

We addressed ethical questions both in **[AE - Comment#2]** and **[R#2 - Comment#1]**. We have reported the number of images and textual content used in each policy's training and testing. Finally, we included Table V to explain how the proposed approach performs in terms of various dimensions of basic needs.

Notice that we were not able to address the large-scale testing comment, i.e., **[R#2 - Comment#2]**. The process of data labelling according to these dimensions is challenging and very time-demanding. Expert psychologists must carefully look at each image in the profile, reach an agreement on label assignment, and fill a common report. It becomes infeasible to annotate a large-scale data set with expert validation.

Reviewer #3 recommended that we (1) include Inter-rater reliability scores and (2) discuss the impact of cluster number and dimensionality of the PCA in the proposed BoC.

We addressed these concerns in responses to **[R#3 - Comment#2]** - **[R#3 - Comment#6]**.

[AE - Comment#2] The manuscript should address the issue of protocol for data collection of profile images, in particular data licensing etc, as raised by Reviewer 2.

Regarding the ethical concerns and data licensing, we did not reflect the whole answers in the manuscript because they are unrelated to the paper's technical aspects. To address this critical concern, we provide answers with the support of ethics and data protection law.

1
2
3 This study complies with applicable international and EU law, specifically EU Directive
4 95/46/E.C. on protecting individuals' privacy and security and the European Union's Charter of
5 Fundamental Rights. When dealing with ethical concerns, the ethical guidelines outlined in the
6 UNESCO Code of Conduct on Social Science Research (P. de Guchteneire, 2006) and the
7 European Code of Conduct on Research Integrity (ESF, 2011) served as guides.
8
9

10 We observed the GDPR's conditions on *Data Erasure* (article 17) and *Privacy by Design* (article
11 23) when developing the research's ethical protocol.
12

13
14 **Right to be Forgotten**

15 Also known as Data Erasure, the right to be forgotten entitles the data subject to have the data
16 controller¹ erase his/her personal data, cease further dissemination of the data, and potentially
17 have third parties halt processing of the data. The conditions for erasure, as outlined in article
18 17, include the data no longer being relevant to original purposes for processing or a data subject
19 withdrawing consent. It should also be noted that this right requires controllers to compare the
20 subjects' rights to "the public interest in the availability of the data" when considering such
21 requests.
22

23
24 **Privacy by Design**

25 Privacy by design as a concept has existed for years, but it is only just becoming part of a legal
26 requirement with the GDPR. At its core, privacy by design calls for the inclusion of data protection
27 from the onset of the designing of systems rather than an addition. More specifically, "The
28 controller shall ... implement appropriate technical and organisational measures ... in an
29 effective way ... in order to meet the requirements of this Regulation and protect the rights of
30 data subjects." Article 23 calls for controllers to hold and process only the data absolutely
31 necessary for the completion of its duties (data minimisation), as well as limiting the access to
32 personal data to those needing to act out the processing.
33

34 For data collection from PUBLIC Instagram profiles, we checked "**Ethics in Social Science and**
35 **Humanities**²," which was published in October 2018, reported on page 9
36

37

38 When processing *social media platform data*:

- 39 • make sure you are sensitive to the issues raised
- 40 • comply with the EU. General Data Protection Regulation (GDPR)
- 41 • *consult your host institution's data protection officer and/or ethics advisor*, and
- 42 • find out if you need to obtain ethical approval for collecting data.

43
44

45
46 The GDPR also provides specific safeguards related to automated processing or profiling of
47 personal data. A panel of experts has drafted these safeguards at the European Commission's
48 request (DG Research and Innovation) on 14 November 2018 as "**Ethics and data protection**³.
49 On page 13 of this guideline, it is stated:
50

51
52

53 ¹ A controller is the entity that determines the purposes, conditions, and means of the processing of personal data.
54 ² https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020_ethics-soc-science-humanities_en.pdf
55 ³ [http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf)
56 [protection_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf)
57
58
59
60

[Box 4] Using 'open source' data

If your research project uses data from *social media networks* and you do not intend to seek the data subjects' *explicit consent* to the use of their data, you must assess whether those persons actually intended to make their information public (e.g. in the light of the privacy settings or limited audience to which the data were made available).

It is not enough that the data be accessible; they must have been made public to the extent that the data subjects do not have any reasonable expectation of privacy. You must also ensure that your intended use of the data complies with any terms and conditions published by the data controller.

If you are in any doubt as to what you can and cannot do with this kind of data, you should seek advice from your DPO or a suitably qualified expert and include their opinion in your proposal.

We also reviewed Instagram's Data Policy to ensure we need to seek the public profiles owner's explicit consent. From Instagram - III. How is this information shared?⁴

Public information can be seen by anyone, on or off our Products, including if they don't have an account. This includes your Instagram username; any information you share with a public audience; information in your public profile on Facebook; and content you share on a Facebook Page, public Instagram account or any other public forum, such as Facebook Marketplace. You, other people using Facebook and Instagram, and we can provide access to or send public information to anyone on or off our Products, including in other Facebook Company Products, in search results, or through tools and APIs. **Public information can also be seen, accessed, reshared or downloaded through third-party services such as search engines, APIs, and offline media such as TV, and by apps, websites and other services that integrate with our Products.**

Given the complexities of the research ethics law for this study, as outlined in the tables above, we sought advice from the UOC Ethics Committee (DPO), which is composed of experts in the field, to design a framework for collecting data from public and private profiles, conducting experiments, and publishing the results. The cases followed in this study are entirely in compliance with their guidelines, which are briefly listed below, and ***we attach the approval obtained from the University Ethics Committee at the end of this answer.***

⁴ <https://help.instagram.com/519522125107875>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

We collected data from 86 public Instagram profiles⁵ (excluding the commercial ones and those of celebrities') by **observing Instagram⁶ and Facebook⁷ data policies**. When gathering data and performing experiments, we adhered to the obligations imposed by Organic Law 3/2018 on Personal Data Protection and Digital Rights, the General Regulation on Data Protection (UE) 2016/679, and the latest complementary legislation. Furthermore, in compliance with the recommendation of the Universitat Oberta de Catalunya's Ethics Committee, **we only reported aggregated results from the experiments (i.e., the average score)**, where we did not ask explicit consent from the owners of the public profiles, to protect the participants' reputation and dignity, as well as to avoid the possibility of disclosing the original users' identities or any personal information from their profiles.

The data management protocol followed was as restrictive as possible. A one-page informative summary of the project had been given to the users while including **only those who returned the signed and written informed consent⁸ in the private group**. Participants were given the option to contact the corresponding author via e-mail to address/resolve any questions or concerns.

For the public group, we used only the data of the users owning public Instagram profiles. To ensure that the user did not change their mind, we downloaded the data while keeping it for just one month. Then we asked two expert psychologists to review **the profiles' visual content**, in which there was no link to personal information, to provide a consensus label for each profile. We trained the proposed pipeline using these labels and **wiped out all data from our secure server once the training was completed. We only kept the links to these profiles**. Therefore, if users delete data or change any settings in the future, our access to this content through links will be affected in compliance with the privacy policy established by each user.

It is the responsibility of all researchers to be concerned about ethical issues. However, a study of the following papers published in reputable journals (including IEEE Transactions on Affective Computing) in which the data is publicly available and each data is associated with a tag or label that links, for instance, the subject to personal traits. In our case, we took a more conservative approach, not making the data publicly available due to ethical considerations, and published the aggregated results.

[1] Teijeiro-Mosquera, L., Biel, J. I., Alba-Castro, J. L., & Gatica-Perez, D. (2014). What your face vlogs about: Expressions of emotion and big-five traits impressions in YouTube. *IEEE Transactions on Affective Computing*, 6(2), 193-205.

⁵ Given that "If your research project uses data from social media networks and you do not intend to seek the data subjects' explicit consent to the use of their data, you must assess whether those persons actually intended to make their information public", **the authors reviewed over 1000 profiles and chose only 76 public profiles that could implicitly cover this requirement**.
⁶ <https://help.instagram.com/519522125107875>
⁷ <https://www.facebook.com/help/203805466323736>
⁸ The consent form was based on the information included in the Horizon 2020 Manual/Ethics "GUIDANCE FOR APPLICANTS-INFORMED CONSENT" published by the European Commission, Research Directorate--General, Directorate L --- Science, Economy, and Society, Unit L3 --- Governance and Ethics. The consent forms were also translated into Persian and given in the written form so that they can be signed.

- [2] Halim, Z., Atif, M., Rashid, A., & Edwin, C. A. (2017). Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits. *IEEE transactions on affective computing*, 10(4), 568-584.
- [3] Biel, J. I., & Gatica-Perez, D. (2012). The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1), 41-55.
- [4] Nobles, A. L., Leas, E. C., Noar, S., Dredze, M., Latkin, C. A., Strathdee, S. A., & Ayers, J. W. (2020). Automated image analysis of instagram posts: Implications for risk perception and communication in public health using a case study of# HIV. *PloS one*, 15(5), e0231155.
- [5] Malighetti, C., Sciara, S., Chirico, A., & Riva, G. (2020). Emotional Expression of# body on Instagram. *Social Media+ Society*, 6(2), 2056305120924771.
- [6] Giordano, G., Primerano, I., & Vitale, P. (2020). A Network-Based Indicator of Travelers Performativity on Instagram. *Social Indicators Research*, 1-19.

Response to Reviewer #1

Recommendation: Author Should Prepare A Minor Revision

Comments:

Good to note that the authors have included an ethical clearance certificate. Good that more data has been collected and new experiments were performed to address the bias issues. Title, keywords and References are revised based on the comments. Good.

We would like to acknowledge you for assisting us with your constructive comments and suggestions in both revisions, which eventually resulted in an increase in the paper's quality.

[R#1 - Comment#1] Out of two annotators, at least one could have been selected who is conversant with Spanish or Iranian. Your multimodal data is collected from Spanish or Iranian. Sometimes certain nuances native to the language can only be interpreted by that particular native language speaker.

Answer: Thank you for pointing out one of the most subtle aspects of this study in the data collection and labelling process. *Linguistic nuances, as you mentioned, are generally understood by native speakers.* In this particular case, we conducted extensive research prior to beginning the data collection process to avoid bias in reporting the findings. We eventually came up with three options, which we will summarise for you.

1. We could concentrate on English language profiles. As detailed in the article, non-English speaking users make up a substantial portion of social network users, and studying their profiles will significantly contribute to the comprehensiveness of artificial intelligence development. On the other hand, such a data set does not exist, and none of us (authors of this paper) is from English-speaking countries. As a consequence, our lack of understanding of the current English-speaking culture could lead to misleading results.
2. Since the psychologists involved in this study speak Persian, another option was to stay focused on Iranian users' Instagram profiles. However, given the connection between basic needs and cultural, social, political, and other factors in the Choice theory, our results were undoubtedly skewed. As a result, we decided on the third option to solve this issue.
3. At this stage, we concentrated on two separate groups of Iranians and Spaniards to present more comprehensive results. The authors of this paper were from Iran and Spain, and distinguishing between business/celebrity and personal profiles to provide a more realistic reflection of society was easier for them. However, the fellow psychologists did not speak Spanish. **As a result, we agreed to exclude the text data from the profiles and only perform the tagging steps by reviewing the visual content.** The visual content frequently evokes common concepts [1], which aids in reducing implicit bias. Nonetheless, we recognised the value of textual content (as shown by the research findings), so we analysed textual content using the proposed algorithm to overcome linguistic limitations. However, the expert labels were given **solely** by visual content; we then used the multimodal data for learning and testing.

We agree with you that a Spanish psychologist's inclusion in this study could be very beneficial in producing more appealing outcomes. However, in response to your comment, we attempted to cover these limitations technically and highlighted this point in several sections of the paper. As an example:

Two expert psychologists (Persian native speakers) who have been trained in Reality Therapy [1] have reviewed **all the visual content of each profile** in the dataset in both phases to provide a consensus annotation per profile. **The primary reason for excluding textual content from the ground-truthing process is that nuances in language are mainly understandable by native language speakers. Nonetheless, visual content often evokes common concepts [2], which helps to minimise implicit bias.** For the multi-label aspect of this study, each profile was labelled with a subset expressed as $L = \{Survival, Belonging, Power, Freedom, Fun\}$, except for the empty set.

[1] Glasser, W. (1999). *Choice theory: A new psychology of personal freedom*. HarperPerennial.

[2] Ware, E. A., Gelman, S. A., & Kleinberg, F. (2013). The medium is the message: Pictures and objects evoke distinct conceptual relations in parent-child conversations. *Merrill-Palmer quarterly (Wayne State University. Press)*, 59(1).

[R#1 - Comment#2] In BoW framework, there will be lot of descriptors. So PCA was handy to reduce the dimension. In the current framework, is PCA necessary? Are there any results without PCA, just for comparison sake at least?

Answer: Thank you for pointing out this comment. In general, an enormous amount of training data is needed in machine learning problems to ensure that there are many samples for each value combination. To prevent the curse of dimensionality, a common rule of thumb is that there should be at least 5 training examples for each dimension in the representation [1]. The curse of dimensionality induces an increase in a classifier's or regressor's average (expected) predictive ability at the cost of degradation rather than steady progress [2-4].

Since we have 86 Instagram profiles, each of them with a 300-D feature vector before applying PCA, If we had not used PCA in the proposed BoC, the occurrence of the curse of dimensionality would have been inevitable. To solve this classic problem in our task, we used PCA to reduce dimensionality. Then, to demonstrate the impact of cluster size and PCA, we conducted detailed analyses, which are available in **Section V.C.2**.

We are aware of your concern and provided the comparative results in this response letter for the sake of comparison. We believe that including this table on paper would alter the logical order in which we have presented our results through various sections and subsections. However, if the respected reviewer suggests that including this table in the manuscript would enhance the paper's technical soundness, we are welcome to do so.

	Architecture (+) PCA	Architecture (-) PCA
Hamming Loss	0.28	0.11
Coverage	3.33	5.81
Ranking Loss	0.35	0.67
Example-based Accuracy	0.68	0.89
F-measure (%)	71.34	48.38

As can be seen in the table, the architecture's performance without PCA is better in measurements relating to each label dimension (e.g., *Hamming Loss* or *Example-based Accuracy*), but it degrades in metrics that can determine the overall performance of the classifier, e.g., *F-measure*.

[1] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[2] Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on pattern analysis and machine intelligence*, (3), 306-307.

[3] Chandrasekaran, B., & Jain, A. K. (1974). Quantisation complexity and independent measurements. *IEEE Transactions on Computers*, 100(1), 102-106.

[4] McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.

Response to Reviewer #2

Recommendation: Author Should Prepare A Major Revision

Comments:

I appreciate the authors' effort for addressing the comments. However, my major concerns still persist:

[R#2 - Comment#1] It is not clear to me whether the people who have public profiles have given their consent, or not. Although it is public, I am not sure it can be used, especially after GDPR, etc. I understand consent has been sought from the people who have private accounts and have seen the psychologist before. However, what about the other participants?

Answer: We fully understand your concern about ethical issues. To address this critical concern, we provide answers with the support of ethics and data protection law.

This study complies with applicable international and EU law, specifically EU Directive 95/46/E.C. on protecting individuals' privacy and security and the European Union's Charter of Fundamental Rights. When dealing with ethical concerns, the ethical guidelines outlined in the UNESCO Code of Conduct on Social Science Research (P. de Guchteneire, 2006) and the European Code of Conduct on Research Integrity (ESF, 2011) served as guides.

We observed the GDPR's conditions on *Data Erasure* (article 17) and *Privacy by Design* (article 23) when developing the research's ethical protocol.

Right to be Forgotten

Also known as Data Erasure, the right to be forgotten entitles the data subject to have the data controller⁹ erase his/her personal data, cease further dissemination of the data, and potentially have third parties halt processing of the data. The conditions for erasure, as outlined in article 17, include the data no longer being relevant to original purposes for processing or a data subject withdrawing consent. It should also be noted that this right requires controllers to compare the subjects' rights to "the public interest in the availability of the data" when considering such requests.

Privacy by Design

Privacy by design as a concept has existed for years, but it is only just becoming part of a legal requirement with the GDPR. At its core, privacy by design calls for the inclusion of data protection from the onset of the designing of systems rather than an addition. More specifically, "The controller shall ... implement appropriate technical and organisational measures ... in an effective way ... in order to meet the requirements of this Regulation and protect the rights of data subjects." Article 23 calls for controllers to hold and process only the data absolutely necessary for the completion of its duties (data minimisation), as well as limiting the access to personal data to those needing to act out the processing.

⁹ A controller is the entity that determines the purposes, conditions, and means of the processing of personal data.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For data collection from PUBLIC Instagram profiles, we checked "**Ethics in Social Science and Humanities**¹⁰," which was published in October 2018, reported on page 9

When processing *social media platform data*:

- make sure you are sensitive to the issues raised
- comply with the EU. General Data Protection Regulation (GDPR)
- *consult your host institution's data protection officer and/or ethics advisor*, and
- find out if you need to obtain ethical approval for collecting data.

The GDPR also provides specific safeguards related to automated processing or profiling of personal data. A panel of experts has drafted these safeguards at the European Commission's request (DG Research and Innovation) on 14 November 2018 as "**Ethics and data protection**¹¹. On page 13 of this guideline, it is stated:

[Box 4] Using 'open source' data

If your research project uses data from *social media networks* and you do not intend to seek the data subjects' *explicit consent* to the use of their data, you must assess whether those persons actually intended to make their information public (e.g. in the light of the privacy settings or limited audience to which the data were made available).

It is not enough that the data be accessible; they must have been made public to the extent that the data subjects do not have any reasonable expectation of privacy. You must also ensure that your intended use of the data complies with any terms and conditions published by the data controller.

If you are in any doubt as to what you can and cannot do with this kind of data, you should seek advice from your DPO or a suitably qualified expert and include their opinion in your proposal.

We also reviewed Instagram's Data Policy to ensure we need to seek the public profiles owner's explicit consent. From Instagram - III. How is this information shared?¹²

Public information can be seen by anyone, on or off our Products, including if they don't have an account. This includes your Instagram username; any information you share with a public audience; information in your public profile on Facebook; and content you share on a Facebook Page, public Instagram account or any other public forum, such as Facebook Marketplace. You, other people using Facebook and Instagram, and we can provide access to or send public information to anyone on or off our Products, including in other Facebook

¹⁰ https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020_ethics-soc-science-humanities_en.pdf
¹¹ http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf
¹² <https://help.instagram.com/519522125107875>

Company Products, in search results, or through tools and APIs. **Public information can also be seen, accessed, reshared or downloaded through third-party services such as search engines, APIs, and offline media such as TV, and by apps, websites and other services that integrate with our Products.**

Given the complexities of the research ethics law for this study, as outlined in the tables above, we sought advice from the UOC Ethics Committee (DPO), which is composed of experts in the field, to design a framework for collecting data from public and private profiles, conducting experiments, and publishing the results. The cases followed in this study are entirely in compliance with their guidelines, which are briefly listed below, and ***we attach the approval obtained from the University Ethics Committee at the end of this answer.***

We collected data from 86 public Instagram profiles¹³ (excluding the commercial ones and those of celebrities') by **observing Instagram¹⁴ and Facebook¹⁵ data policies**. When gathering data and performing experiments, we adhered to the obligations imposed by Organic Law 3/2018 on Personal Data Protection and Digital Rights, the General Regulation on Data Protection (UE) 2016/679, and the latest complementary legislation. Furthermore, in compliance with the recommendation of the Universitat Oberta de Catalunya's Ethics Committee, **we only reported aggregated results from the experiments (i.e., the average score)**, where we did not ask explicit consent from the owners of the public profiles, to protect the participants' reputation and dignity, as well as to avoid the possibility of disclosing the original users' identities or any personal information from their profiles.

The data management protocol followed was as restrictive as possible. A one-page informative summary of the project had been given to the users while including **only those who returned the signed and written informed consent¹⁶ in the private group**. Participants were given the option to contact the corresponding author via e-mail to address/resolve any questions or concerns.

For the public group, we used only the data of the users owning public Instagram profiles. To ensure that the user did not change their mind, we downloaded the data while keeping it for just one month. Then we asked two expert psychologists to review **the profiles' visual content**, in which there was no link to personal information, to provide a consensus label for each profile. We trained the proposed pipeline using these labels and **wiped out all data from our secure server once the training was completed. We only kept the links to these profiles.**

¹³ Given that "If your research project uses data from social media networks and you do not intend to seek the data subjects' explicit consent to the use of their data, you must assess whether those persons actually intended to make their information public", **the authors reviewed over 1000 profiles and chose only 76 public profiles that could implicitly cover this requirement.**

¹⁴ <https://help.instagram.com/519522125107875>

¹⁵ <https://www.facebook.com/help/203805466323736>

¹⁶ The consent form was based on the information included in the Horizon 2020 Manual/Ethics "GUIDANCE FOR APPLICANTS-INFORMED CONSENT" published by the European Commission, Research Directorate--General, Directorate L --- Science, Economy, and Society, Unit L3 --- Governance and Ethics. The consent forms were also translated into Persian and given in the written form so that they can be signed.

Therefore, if users delete data or change any settings in the future, our access to this content through links will be affected in compliance with the privacy policy established by each user.

It is the responsibility of all researchers to be concerned about ethical issues. However, a study of the following papers published in reputable journals (including IEEE Transactions on Affective Computing) in which the data is publicly available and each data is associated with a tag or label that links, for instance, the subject to personal traits. In our case, we took a more conservative approach, not making the data publicly available due to ethical considerations, and published the aggregated results.

- [1] Teijeiro-Mosquera, L., Biel, J. I., Alba-Castro, J. L., & Gatica-Perez, D. (2014). What your face vlogs about: Expressions of emotion and big-five traits impressions in YouTube. *IEEE Transactions on Affective Computing*, 6(2), 193-205.
- [2] Halim, Z., Atif, M., Rashid, A., & Edwin, C. A. (2017). Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits. *IEEE transactions on affective computing*, 10(4), 568-584.
- [3] Biel, J. I., & Gatica-Perez, D. (2012). The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1), 41-55.
- [4] Nobles, A. L., Leas, E. C., Noar, S., Dredze, M., Latkin, C. A., Strathdee, S. A., & Ayers, J. W. (2020). Automated image analysis of instagram posts: Implications for risk perception and communication in public health using a case study of# HIV. *PloS one*, 15(5), e0231155.
- [5] Malighetti, C., Sciara, S., Chirico, A., & Riva, G. (2020). Emotional Expression of# body on Instagram. *Social Media+ Society*, 6(2), 2056305120924771.
- [6] Giordano, G., Primerano, I., & Vitale, P. (2020). A Network-Based Indicator of Travelers Performativity on Instagram. *Social Indicators Research*, 1-19.

[R#2 - Comment#2] Although the paper is interesting from a technical viewpoint, the dataset is very biased and limited – 86 profiles only. Private group has 10 profiles. I fully understand collecting data from private group is not a trivial task. However, I think first the method should be validated on general profiles at a large scale and then should be tested on this specific population.

Answer: Thank you for your constructive feedback. We understand your concern because we had the same one before starting this research. However, not only is it challenging to collect data from a private group, but it is also hard for expert psychologists to label profiles. As we explained in our response to **[R#2 - Comment#4]**, 37,530 feeds containing 38,356 photos were checked by expert psychologists. These feeds had to be carefully examined in order to find the best combination of labels that could describe the basic needs of these 86 profiles. **Reviewing each profile took about three to four hours for both psychologists**, and this process must be repeated for each phase of data collection. *Furthermore, there is no large, publicly accessible data set with labels for the five basic needs.* As a result, we took a number of steps in this study to reduce unintentional bias:

- Are users revealing their **true needs** and mental states, or they are **trying to build** a perfect image through their social media profiles to escape from real life? *This concern has always existed in the field of affective computing*, where for example, we trust the available tagged data in the training of an emotional recognition model, *apart from the fact that the image may represent true or imposed happiness*. Glasser's choice theory [1] was built on the basis of our choices when **we chose to show or pretend happiness**.

That is why this study focused on choice theory to alleviate bias from this viewpoint and open up another dimension to affective computing.

- Since the sample size (86 Instagram profiles) is far less than the total number of Instagram accounts (over one billion active users, an unintentional sampling bias could occur. Following the suggestion in [3], we hypothesised that if the distribution of basic needs for Iranian and Spanish users differs significantly, **the probability of sampling bias is insignificant**. We addressed your concern regarding this hypothesis in [R#2 - Comment#3].
- In the third policy, **we first examined the bias effect and generalisability of the proposed approach** for predicting labels from new profiles. We also evaluated the dependency of the proposed approach on the size of the data. To analyse the bias, we train the proposed architecture with data collected in Phase 1 and test it with data collected in Phase 2 (see Table IV). To analyse the scale dependence, we used the cross-validation strategy (leave one subject out) only for the portion of the data collected in Phase 2. The results are shown in Table V.

Furthermore, as discussed on page 580 of "The SAGE Handbook of Social Media Research Methods - Chapter 34" [3], under the title "Small Samples of Instagram Data,"

“Small data approaches relying primarily on qualitative analysis offer 'a granularity of detail that might otherwise be lost in dazzling large-scale data visualisations that value the quantitative over the qualitative' (Losh, 2015: 1650). As this 'era of Big Data' has contributed to discourse that discounts the value of qualitative research and small sample sizes (boyd and Crawford, 2012), the value of these approaches to Instagram data warrants being stated explicitly. Researchers looking to make use of rich data generally face a tradeoff between depth and breadth, with an inverse relationship between the amount of usable data gathered from each post or Instagram user and the size of their sample (Morse, 2000). Accordingly, the analysis of small samples of Instagram data can provide extremely valuable insights that could not be obtained from Big Data approaches.”

[1] Glasser, W. (1999). *Choice theory: A new psychology of personal freedom*. HarperPerennial.

[2] Ware, E. A., Gelman, S. A., & Kleinberg, F. (2013). The medium is the message: Pictures and objects evoke distinct conceptual relations in parent-child conversations. *Merrill-Palmer quarterly (Wayne State University. Press)*, 59(1).

[3] Sloan, L., & Quan-Haase, A. (Eds.). (2017). *The SAGE handbook of social media research methods*. Sage.

[R#2 - Comment#3] Also, the authors have hypothesised that if the distribution of basic needs for Iranian and Spanish users differs significantly, the probability of sampling bias is insignificant. It is not clear why this should be the case.

Answer: According to Glasser's choice theory [1], human choices are influenced by several factors, including environmental factors, cultural, economic, and political circumstances. In this study, we first reviewed over 1000 Instagram profiles before narrowing it down to 86 based on our knowledge of Iran and Spain's social, political, economic, and cultural conditions. However, we needed to ensure the diversity of the basic needs distribution across these profiles before labelling and conducting the rest of the analysis. Furthermore, we must be certain that the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

classifier will not overfit on a particular dimension of needs (e.g., love or survival) as a result of the underlying distribution of needs in the selected profiles. In order to statistically endorse our work, we hypothesised that if the distribution of basic needs for Iranian and Spanish users differs significantly, the probability of sampling bias is insignificant. This hypothesis was motivated by the methodology presented in [2].

We hypothesise that the reasons for the difference in need distributions emerge from political, economic, and social factors that vary significantly between the two countries. Nonetheless, knowing the fundamental motives of both populations is beyond the scope of this study. The provided diversity enriched the learning task.

[1] Glasser, W. (1999). *Choice theory: A new psychology of personal freedom*. HarperPerennial.
[2] Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

[R#2 - Comment#4] How many images have been used for training and testing in the end? It is not clear how many images have been used in each policy.

Answer: Thank you for bringing this up. Although we listed the exact number of collected feeds (visual and textual contents) in the first version of the paper, we removed it from the revision file at the UOC ethical committee's request, which asked us only to report aggregated results. However, we agree that stating this statistic does not compromise research ethics. As a result, we addressed your concern in the following manner:

Section III - Data Collection: *In total, we collected 30,080 feeds (each feed may have multiple images) in the first phase and 7,450 feeds in the second phase.*

Section VA - Subjective Test: *In this test, eight bilingual volunteers were asked to annotate the visual content of two Instagram profiles, each with an average of 286 feeds, using the choice theory.*

Section VC - Experiments: *We used three policies to train and evaluate the GLOCAL classifier in our experiments. In total, 38,356 images were sent to image descriptors, and 6,943,050 words were used to refine the FastText model. In the first policy, We used the cross-validation strategy (leave one subject out) to handle the small sample size. To minimise statistical uncertainty, the results were averaged across independent repetitions. In the second policy, we trained GLOCAL with 77 samples (35,382 images and 6,619,614 words) and validated it with the nine samples used for the subjective test. This 90/10 split helps us to compare the proposed approach to the subjective test. In the third policy, we trained GLOCAL with all of the samples collected in Phase 1 (30,824 images and 5,594,880 words) and tested it with samples (7,532 images and 1,348,170 words) added to the dataset in Phase 2.*

[Comment#5] The results have not been presented with respect to different dimensions (e.g., survival, belonging, etc.), it is not clear how the method is performing for different profiles, especially considering the uneven distribution.

Answer: We would like to thank you for bringing this to our attention. In the third policy, we trained GLOCAL with all of the samples collected in Phase 1 and tested it with samples added

to the dataset in Phase 2. In fact, in this policy, we examined how the proposed approach would behave in the presence of the uneven and evolving distribution of needs. We reported results for different dimensions of the basic needs in Table V.

Table V - Performance of the proposed approach considering class-wise recall, precision, and F1 score. Here, TP, TN, FP and FN stand for True Positive, True Negative, False Positive and False Negative, respectively.

	TP	TN	FP	FN	Precision	Recall	F1-score
Survival	29	29	10	18	0.74	0.62	0.67
Belonging	17	55	9	5	0.65	0.77	0.71
Power	12	45	15	14	0.44	0.46	0.45
Freedom	10	48	18	10	0.36	0.5	0.42
Fun	62	10	7	7	0.9	0.9	0.9

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Response to Reviewer #3

Recommendation: Author Should Prepare A Minor Revision

[R#3 - Comment#1] Authors mention that "In this study, we used the choice theory to assess mental health from..." Authors relate Glasser's choice theory to assess mental health. However, this is not what is performed in this paper. The goal is to categorise user profiles in terms of five basic needs, no related mental health assessment was performed.

Answer: Thank you for pointing out this issue. We rephrased in the paper as

This study used the choice theory to categorise users' profiles considering five basic needs from a broader yet language-independent perspective.

[R#3 - Comment#2] Inter-rater reliability scores should be added.

Answer: The inter-rater reliability scores can be measured for ground-truthing procedure and subjective test.

Since two psychologists reviewed the Instagram profiles and provided a consensus label for each profile, measuring this score for consensus opinion is infeasible.

In the subjective test, eight bilingual volunteers were asked to review the visual content of two Instagram profiles and annotate them according to the choice theory. We divided the participants into two groups with a gender ratio of 1. Four Persian/English speakers and four Spanish/English speakers were assigned to $G_{Persian}$ and $G_{Spanish}$, respectively. We also randomly selected four public profiles belonging to Iranian users and four public profiles belonging to Spanish users from our data set and added them to TG_{Iran} and TG_{Spain} , respectively.

We asked all of the volunteers to review the Instagram profile of the main author. Since all volunteers have known the main author personally, the review of this profile was somewhat similar to the assessment of a profile by expert psychologists when an in-person diagnosis is available. We also asked $G_{Persian}$ members to visit TG_{Spain} profiles and $G_{Spanish}$ members to review TG_{Iran} profiles. We have done this to avoid unintentional bias to language and personal bonding. The results of the subjective test are shown in Table I.

We assess the consistency of the non-expert and expert observers' quantitative measurements by measuring the **intra-class correlation coefficient** (ICC) [1]. We chose the "A-1" type to measure ICC because we were interested in measuring the absolute agreement between the two raters in the presence of random residual errors and the two raters' systemic errors. The mean intra-class correlation coefficient and confidence levels are $\underline{r} = 0.22$ and $\underline{p} = 0.64$, respectively, implying a *low* intra-class correlation.

[1] McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), 30.

[R#3 - Comment#3] It is not clear why the authors chose items in the densest cluster (instead of the items that are closest to each cluster centroid) while obtaining textual content representation. Please clarify.

Answer: As we explained in the paper, the overall routine is as follows:

Suppose that each feed contains M words, the representative matrix for this feed has the size of $M \times 300$. We then apply k -means [1] with $k = 5$ (analogous to 5 basic needs) to rank this matrix. The items belonging to the densest cluster are stored in a list to select the most informative rows.

In cluster analysis, there are three common ways to select representative cluster(s): (1) groups with small distances between cluster members (as you stated in your comment), (2) dense areas of the data space, and (3) intervals or specific statistical distributions. We chose the densest cluster for our study because objects eventually converge to the local maxima of density, and the densest cluster can thus serve as a representative for the data set. This impact has been thoroughly investigated in [2].

[1] Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding*. Stanford.

[2] Achtert, E., Böhm, C., & Kröger, P. (2006, April). DeLi-Clu: boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 119-128). Springer, Berlin, Heidelberg.

[R#3 - Comment#4] BoC description needs to be improved. After codebook (cluster centroids) are identified, each word should be assigned to a cluster and then occurrence/presence of words in all clusters should denote the BoC representation (PCA may follow for further dimensionality reduction). This is not clear in Algorithm 1.

Answer: In the proposed BoC, we first concatenate the high-level descriptors $S_i, i = \{1,2,3\}$ and then use the fine-tuned FastText word embedding model to transform the list of tokenised documents to sequences of numerical vectors. The vectors are then clustered into λ clusters using the k -means algorithm, with the cluster centres serving as the representative dictionary. We used PCA [28] to map the $\lambda \times 300$ representative dictionary into a d -dimensional feature vector, where $d \ll \lambda$ is the number of dimensions. This low-dimensional representation can handle the small sample size, the data set's imbalanced characteristics, and the possibility of the curse of dimensionality when training the classifier.

[R#3 - Comment#5] Do the authors train BoC using a subset of profiles? Please clarify.

Answer: The BoC consists of three modules. FastText, k -means, and PCA. We fine-tuned the FastText model with all terms in our data and with the category labels in Places2 and Microsoft COCO datasets. For the k -means and PCA, which are sorts of clustering algorithms, we used all terms in S_1, S_2 , and S_3 which are representative features of all profiles.

[R#3 - Comment#6] It is not clear why the authors have chosen d so small (in the set $\{1,2,3\}$). It looks like $d = 2$ cannot explain at least 95% of the variance.

Answer: Thank you for bringing this to our attention. In fact, the PCA dimensions (d) are not the only factors influencing the classifier's performance. Indeed, both the number of clusters (λ) and the PCA dimensions (d) affect the efficiency of the proposed approach. While $d = 2$ cannot explain at least 95% of the variance, we discovered in our experiments that $\{\lambda, d\} = \{30, 2\}$ could result in the best performance for the multi-label classification task in our study. To clarify, we rewrite **Section V.C.2** as follows:

The number of clusters (λ) and PCA dimensions (d) contribute to the performance of the proposed method. We first plotted the number of terms in the training data (see Fig. 7) and found that the distribution of BoC terms is within [30, 70].

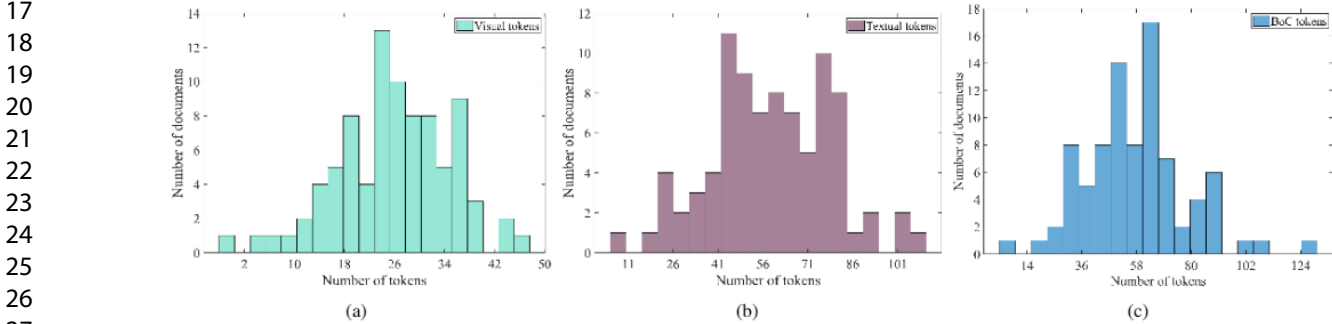


Fig. 7. Terms count histogram for (a) visual-extracted terms, (b) textual-extracted terms, and (c) BoC-extracted terms.

Then to find the most appropriate cluster number (λ) and to analyse the impact of applying PCA to the representative dictionary, we trained GLOCAL with the first policy as a function of (λ, d) and plotted the Hamming Loss. Figure 8 shows the results.

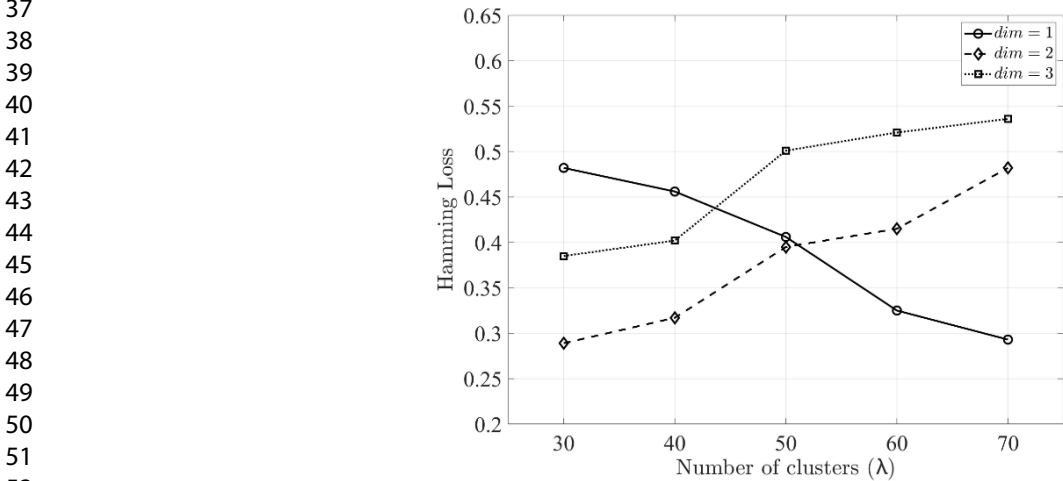


Fig. 8. Hamming loss of GLOCAL as a function of (λ, d). GLOCAL achieved HL = 0.293 with $\{\lambda, d\} = \{70, 1\}$ and HL = 0.289 with $\{\lambda, d\} = \{30, 2\}$. In $\{\lambda, d\} = \{30, 2\}$ vectors were concatenated before passing to the classifier.

Figure 8 shows that the minimum values of hamming loss (with a minor difference) obtained for the two combinations of $\{\lambda, d\} = \{30, 2\}$ and $\{\lambda, d\} = \{70, 1\}$. We measured the variance of the principal component in order to select the appropriate d , where $\lambda = \{30, 70\}$. We found that $d = 2$ is capable of explaining $\{89.6\%, 49.6\%\}$ of the total variances in relation to the corresponding λ .