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

Mohammad Mahdi Dehshibi[1], Bita Baiani[2], Gerard Pons[3], David Masip[1]
[1]Department of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
[2]Department of Psychology, Islamic Azad University, Science and Research Branch, Tehran, Iran
[3]Centrum 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 through 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 processes of preventive measures, 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 adapted[1] 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, 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 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 is not only aligned with three out of 10 axioms of Glasser's choice theory[2] but it 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 and language-

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

[1]Cultural 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.

independent perspective. To this end, we have studied how people implicitly contribute to unmet basic needs by *choosing* contents to share on their Instagram profile in order to draw a personal picture of their quality worlds. We also explored how visual and textual contents can tell us about (1) those about whom we care and those who care about us (Belonging); (2) those we respect and those 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 a year and a half, we collected data from 86 public Instagram profiles (excluding business, influencers and celebrities) while observing Instagram[3] and Facebook[4] 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 subnetwork 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 results. In Section VI, we discuss the limitations and challenges of this work. Finally, Section VII concludes the paper.

[3]https://help.instagram.com/519522125107875

[4]https://www.facebook.com/help/203805466323736

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 sharing photos is more straightforward than by providing texts, 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 social withdrawal, emotional-context insensitivity, reduced reactivity hypotheses of depression, gender dimension and 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 assess mental health from a broader and 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.

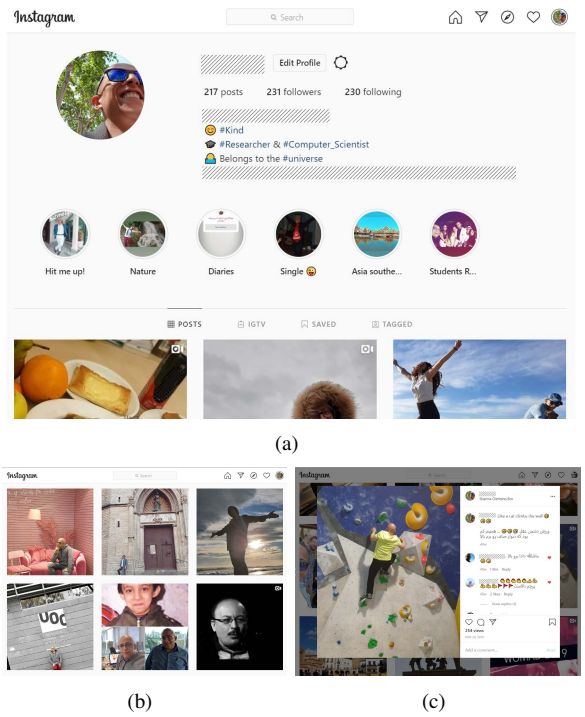


Fig. 1. A sample Instagram profile with (a) bio, (b) feeds and (c) post, including caption and comments. The textual content of this sample is in English and Persian, including emojis, hashtags and geotag. Please note that personal information has been masked for the purpose of publishing and observing 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 research purpose and about how the data would be used. Finally, ten potential participants (5 males and 5 females) who were fully informed and had 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. Having considered the recommendation of the lead psychologist and our previous experience in data collection [30, 31, 32, 33], 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 lead 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 contents of each profile in the dataset in both phases to provide a consensus annotation for each profile. The primary reason for excluding textual contents from the ground-truthing process is that nuances in language are mainly understandable by native language speakers. Nonetheless, visual contents often evoke common concepts [11], which help 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 Fig. 2.

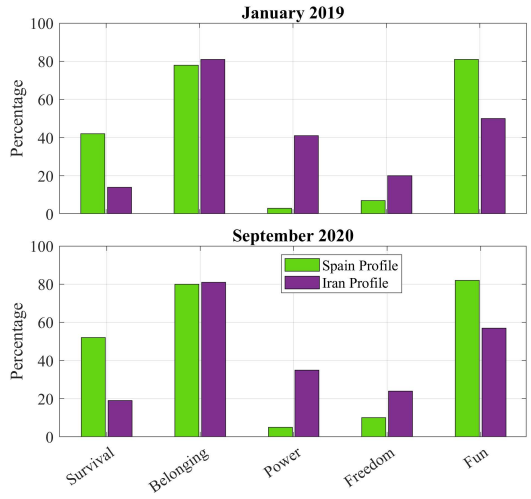


Fig. 2. Diversity of labels per country. The top row shows the percentage of each basic need in January 2019 and the bottom row shows these statistics in September 2020.

In the data acquisition, we used random sampling to ensure that our samples could be representatives of the population. Since the sample size (86 Instagram profiles) is far less than

the total number of Instagram accounts (over one billion active users[5]), an unintentional sampling bias could occur. Following the suggestion in [34], 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 [35, 36] 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 which the following two can be mentioned: 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 perception 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), it 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 was given to the users, while including only those who had returned the signed and written informed consent[6] in the private group. The 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 contents of the profiles, in which there was no link to personal information, to provide a consensus label for each profile. We trained the proposed architecture using these labels and wiped out all data from our secure server once the training was completed. We only kept the links to those profiles. Therefore, if users delete data or change any settings in the future, our access to these contents 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 the participants, including the information sheet, and the procedure for the recruitment of the subjects, the committee approved 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

A users can use both visual and textual contents on their Instagram profile to create a personal photo album. While the owner of the profile can exclusively use the visual contents, the textual contents 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 the psychologists to perceive the basic needs of the users 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 contents to strengthen the representation of their feelings on their profile. Thanks to significant advances in natural language processing, in this research, we could address linguistic challenges and use both modalities in the assessment of the Instagram profiles.

We proposed a multimodal and multi-label classification for perceiving the five basic needs in accordance with Glasser's choice theory. For the representation of the visual contents, we use the Places-CNN [37] scene descriptor and the YOLO-based object detector [21], which have been trained with the Places2 [22] and the Microsoft COCO datasets [23], respectively. To represent textual contents, 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 sub-sections. 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. We use Places-CNN [37] to estimate the probabilities of 365 categories of places, such as 'lake natural', 'restaurant', 'downtown' and 'train station platform'. 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

[5]https://www.statista.com/topics/1882/instagram/

[6]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.

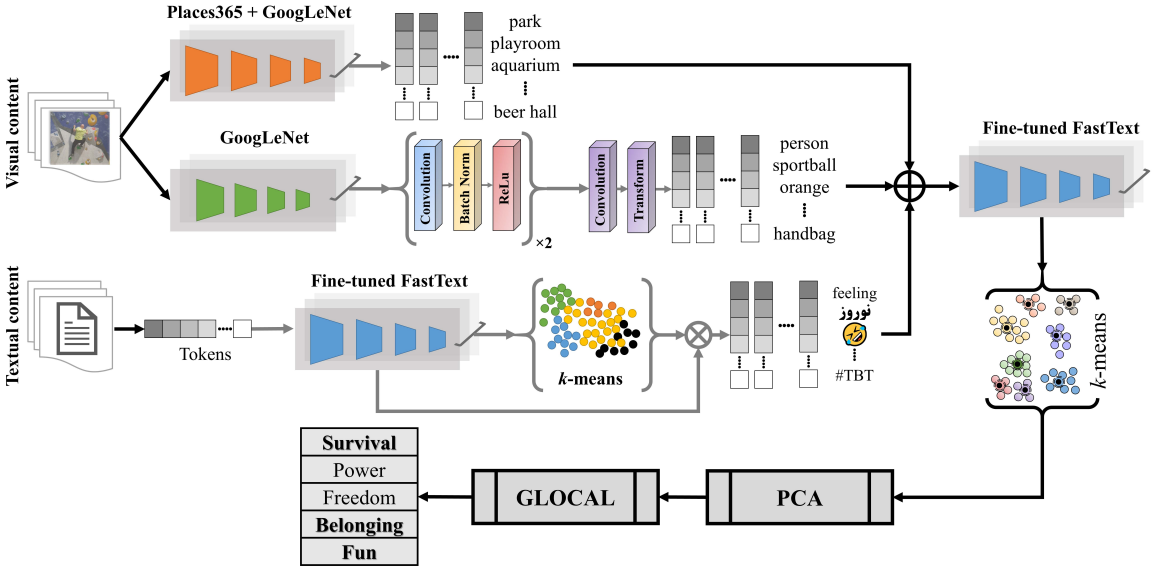
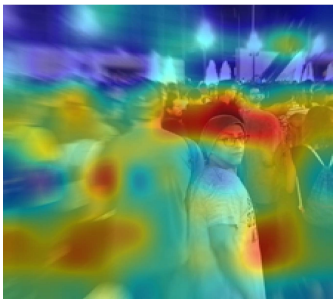


Fig. 3. The pipeline of the proposed approach to perceive the basic needs of the Instagram users. We used Places-CNN [22] and the modified YOLO-based object detector [21], which were trained with the Places2 [22] and the Microsoft COCO datasets [23], respectively, to extract places and objects. We fine-tuned 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 $\bigoplus$ to show the appending of all words together, $\bigotimes$ to show the numerical vector mapping to words, and added $\bullet$ inside each cluster centre.

were concatenated to represent a profile with a list of scenes (S_1). Note that for each profile, the S_1 size is different and may include duplicated place tags. Two examples of scene descriptions are given in Fig. 4.



1: restaurant kitchen
2: beer hall
⋮
19: swimming hole

(a)

1: sandbox
2: playroom
⋮
10: aquarium

(b)

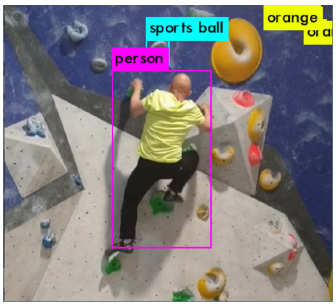
Fig. 4. Places-CNN with GoogLeNet backbone [37] trained with Places2 dataset [22] is used to classify scenes in images. Only categories with a probability of 0.01 or higher were considered. (a) An outdoor and (b) indoor activity with potential tags like 'restaurant kitchen', 'beer hall' and 'playroom'. 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 in each 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.

1: person
2: person
:
14: person
15: handbag

(a)



1: person
2: sport ball
3: orange
4: orange

(b)

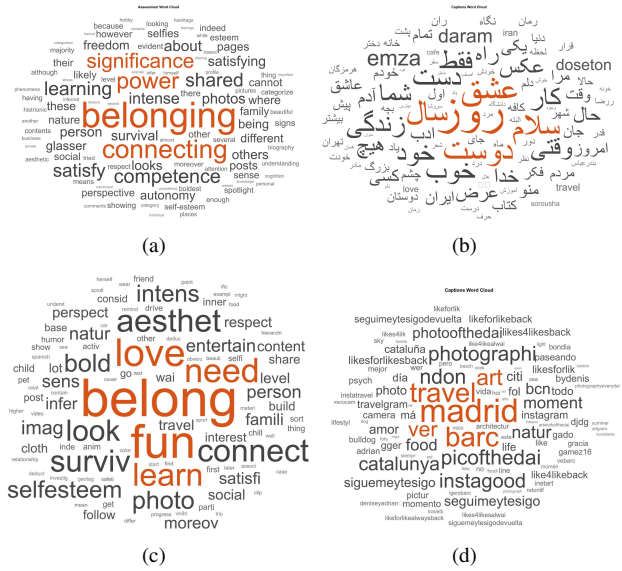
Fig. 5. The YOLO-based object detector with the modified detection sub-network is used to extract the probability and bounding boxes of the COCO dataset categories [23]. (a) Participation in an outdoor activity in which 14 objects are detected as 'person' and one object as 'handbag'. (b) Indoor climbing with four detected objects.

B. Textual content representation

As we explained, in order to the tag Instagram profiles for the five basic needs, the expert psychologists have reviewed only the visual contents of the profiles to overcome language barriers and not to be overwhelmed by the contents of the texts written in their native language. However, not only has our previous experience [38, 39, 40] but also studies targeting the self-disclosure behaviour of social media users [18, 19, 20] have shown that textual contents are important cues for assessing users' attitudes, needs, and mental states.

To understand the importance of textual contents, we represent the perceptions of the psychologists as well as the texts written by the users in the two word clouds, see Fig. 6. The word cloud records the cumulative frequency of words in each document. Although we were concerned about the domain intersection due to the translation of the psychologists' perceptions from Persian into English and the multilingual nature of the texts used by the users (sometimes in a feed), the observed semantic association inspired us to add the textual modality to enrich the latent space.

We use the FastText model [24] to create a textual embedding representation. In the FastText, words are broken into several sub-words (n-grams), and the word embedding vectors will be the sum of all these n-grams. The pre-trained FastText model with Wikipedia data contains most standard words in languages. However, representing creative words often appearing in a multilingual form and/or hashtag required us to fine-tune the FastText with all terms in our dataset as well as the category labels in Places2 and Microsoft COCO datasets.



(a) (b) (c) (d)

Fig. 6. The word clouds related to (a, c) the psychologists' perceptions, (b, d) the textual contents of the profiles. [Top] Iran profiles, where {day, love, eye, friend, window} (translated into English) are the top 5 words used by Iranian profile owners and {belonging, page, love, connect, satisfy} are the top 5 words outlined by the expert psychologists in reviewing the Iranian profiles. [Bottom] Spain profiles, in which the top 5 words used by the users are {madrid, barc, travel, art, ver}, and the top 5 words used by the psychologists to outline their perceptions are {belong, fun, love, need, learn}.

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 [41] 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 contents 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 ones. To harmonise these three representations, we use *vec2word* tool [42] to convert the numerical vectors of S_3 into the categorical one.

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

In this research, we used high-level descriptors to map both the textual and visual contents of the profile into the three sets of terms, i.e., S_1, S_2, S_3. Since each list has a different cardinality, a uniform codebook cannot be constructed by concatenating these three lists. In addition, the calculation of the terms' frequency does not necessarily guarantee the best representation since the relevance and importance of the terms are not recorded; for instance, see Figures 4 and 5.

We proposed *Bag-of-Content* to build a codebook that represents both visual and textual modalities of a profile to train 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 a representative dictionary. We used Principal Component Analysis (PCA) [43] to map the $\lambda \times 300$ representative dictionary into a d-dimensional feature vector where $d << \lambda$ is the number of dimensions. This low-dimensional representation can handle the small sample size, the dataset'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.

$\mathcal{P} \leftarrow \bigcup_{i=1}^{3} S_i$;
$n = |\mathcal{P}|$;
// n: Number of elements in $\mathcal{P}$.
$\mathcal{V}_{n\times 300} \leftarrow$ FastText$(\mathcal{P})$;
// $\mathcal{V}$: Numerical representation of $\mathcal{P}$
Apply k-means$(\mathcal{V}_{n\times 300},\ \lambda)$ and add cluster centres to $codebook_{\lambda\times 300}$;
if $d > \min(n,\ 300)$ **then**
    $d = \min(n,\ 300)$;
end
$\text{BoC}_{\lambda\times d} \leftarrow$ PCA$(codebook_{\lambda\times 300}, d)$
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 an overlap or a conflict with the basic needs of a 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 the both global and local label correlations.

Let the set of l class labels be expressed in the form of $C = \{c_1, \cdots, c_l\}$. 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 uses the regularisation of the global and local manifolds as well as low-rank decomposition to utilise the both global and local label correlations. The label matrix in GLOCAL is decomposed into 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 ($\tilde{Y}$).

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 those of 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 contents 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 dataset 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 the expert psychologists when an in-person diagnosis is available. The results of the subjective test are shown in Table I.

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

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

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 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 calculating the intraclass correlation coefficient (ICC) [44]. We chose the 'A-1' type to calculate 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 of intraclass correlation coefficient and that of confidence level are $\tilde{r} = 0.22$ and $\tilde{p} = 0.64$, respectively, implying a *low* intraclass correlation. Contrary to the suggestion made by [45, 46], these statistics also indicate that the annotation of data on this specific topic – requiring psychological expertise – cannot benefit from the recruitment of workers from crowdsourcing 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 the GoogLeNet backbone [37], 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* [42]. We also set the rejection threshold to 10^{-4} to sub-sample the most frequent words (see [42] for more details). We downloaded Wikipedia dumps of Spanish, Persian, English, and Turkish languages. We normalised the raw Wikipedia data using a semantic vector tool[7] 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 $\mathrm{HL}(H, X) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \triangle \tilde{Y}_i|}{|C|}$, where H is the generated model by the multi-label classifier, N is the number of test data, and $\triangle$ 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 of the example. It can be calculated by $\mathrm{Cvg}(H, X) = \frac{1}{N} \sum_{i=1}^{N} \max(r_i(c)) - 1$, 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).

[7] https://github.com/PrincetonML/SemanticVector

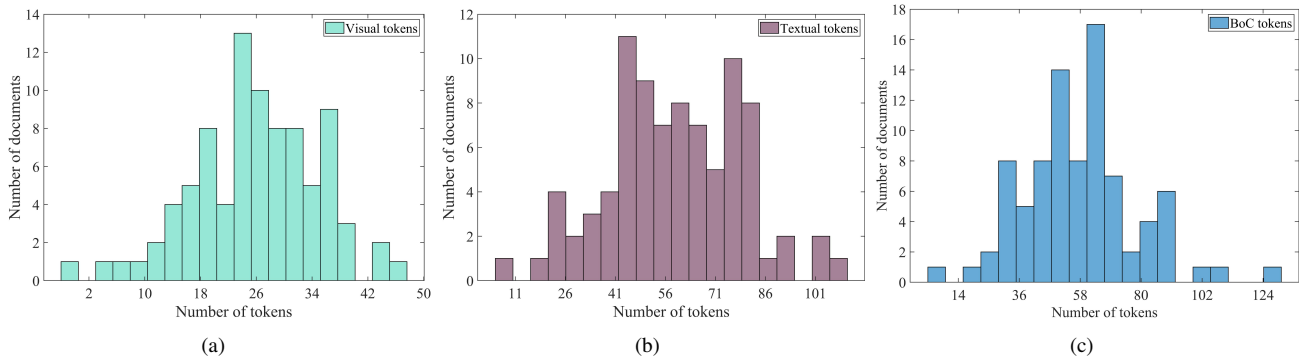


Fig. 7. Terms count histogram for (a) visual-extracted terms, (b), textual-extracted terms, and (c) BoC-extracted terms.

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

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.

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

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 Fig. 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 Fig. 9), where vectors were concatenated before passing to the classifier.

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

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

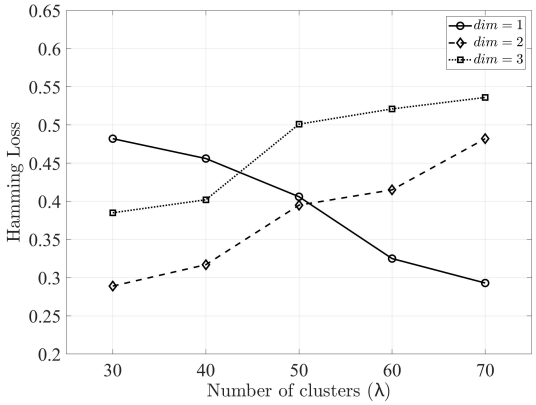


Fig. 8. Hamming loss of GLOCAL as a function of $\{\lambda, d\}$. GLOCAL achieved HL $= 0.293$ with $\{\lambda, d\} = \{70, 1\}$ and HL $= 0.289$ with $\{\lambda, d\} = \{30, 2\}$. In $\{\lambda, d\} = \{30, 2\}$ vectors were concatenated before passing to the classifier.

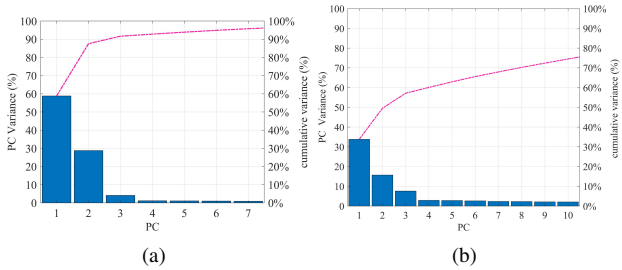


Fig. 9. Pareto plot for variance percentages of the PCA dimensions, where (a) $\lambda = 30$ and (b) $\lambda = 70$. In this plot, PC stands for the principal component.

TABLE II
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

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 Fig. 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

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, $\Upsilon^{Places}_{86\times365}$ and $\Upsilon^{YOLO}_{86\times80}$ 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, $\Upsilon^{FastText}_{86\times300}$ 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 $\Upsilon^{Fusion}_{86\times86}$ 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.

VI. DISCUSSION

Perceiving human basic needs from Instagram profiles was associated with its own set of limitations and challenges. We collected multimodal data from Spanish and Iranian Instagram

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

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. $\ominus$ 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	60	0.28 ± 0.03	3.33 ± 0.23	0.35 ± 0.05	0.68 ± 0.02	71.34 ± 1.64
Proposed $\ominus$ YOLO	60	0.32 ± 0.08	3.77 ± 0.11	0.37 ± 0.10	0.59 ± 0.08	68.94 ± 2.10
Proposed $\ominus$ 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

profiles, and we asked our Persian fellow psychologists trained in Reality Therapy [13], who did not speak Spanish, to provide us with the ground truth. Sometimes certain nuances native to the language can only be interpreted by that particular native language speaker, which ignoring it can introduce unintended bias in reporting the results. Accordingly, we resolved linguistic restrictions in our experimental protocol by (1) excluding text data from the profiles and asking fellow psychologists to only perform tagging steps by evaluating visual content and (2) analysing textual content using the proposed algorithm. However, we agree that having a Spanish psychologist fellow, and in general, native expert annotators in these types of studies, could be extremely beneficial in reporting more appealing results.

Another challenge of this study was the limited number of our dataset, which contained only 86 profiles, ten of which belonged to the private group. However, it should be mentioned that it is not only difficult to collect data from a private group, but it is also difficult for expert psychologists to label profiles. As we explained in Section III, expert psychologists reviewed 37,530 feeds containing 38,356 images. These feeds had to be carefully analysed to find the best label combination that could describe the basic needs of these 86 profiles. For both psychologists, reviewing each profile took about three to four hours, and this procedure 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. Since the sample size is much less than the total number of Instagram accounts (over one billion active users), we statistically endorse our work by hypothesising that if the distribution of basic needs differs significantly between Iranian and Spanish users, the probability of sampling bias is insignificant. The explanation for this hypothesis comes from the fact that differences in need distributions are caused by political, economic, and social factors that differ greatly between the two countries. However, understanding the underlying motivations of both populations is beyond the scope of this research.

Finally, we must answer an important question: "Are users sharing their true needs and mental states, or are they attempting to create an idealised picture through their social media profiles to escape from real life?" This concern has always existed in the field of affective computing, where we rely on available tagged data to train an emotional recognition model, despite the fact that the image can represent true or imposed emotion. Glasser's choice theory [13] was based on humans' choices to disclose or pretend their true needs. As a result, the focus of this study was on choice theory to eliminate bias from this viewpoint and open up another dimension to affective computing.

VII. 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, "Vicsom: 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: A structural equation modeling analysis," *Computers in Human Behavior*, vol. 111, p. 106388, 2020.

[4] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017.

[5] D. Muriello, L. Donahue, D. Ben-David, U. Ozertem, and R. Shilon, "Under the hood: Suicide prevention tools powered by ai," *Facebook Code*, 2018.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.

[7] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[8] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches." in *ICWSM*, 2008.

[9] P. Cuijpers, O. Eylem, E. Karyotaki, X. Zhou, and M. Sijbrandij, "Psychotherapy for depression and anxiety in low-and middle-income countries," in *Global Mental Health and Psychotherapy*. Elsevier, 2019, pp. 173–192.

[10] G. Bernal, J. Bonilla, and C. Bellido, "Ecological validity and cultural sensitivity for outcome research: Issues for the cultural adaptation and development of psychosocial treatments with hispanics," *Journal of abnormal child psychology*, vol. 23, no. 1, pp. 67–82, 1995.

[11] E. A. Ware, S. A. Gelman, and F. Kleinberg, "The medium is the message: Pictures and objects evoke distinct conceptual relations in parent-child conversations," *Merrill-Palmer quarterly (Wayne State University. Press)*, vol. 59, no. 1, 2013.

[12] A. T. Beck, *Depression: Clinical, experimental, and theoretical aspects*. Hoeber Medical Division, Harper & Row, 1967.

[13] W. Glasser, *Choice theory: A new psychology of personal freedom.* HarperCollins Publishers., 1998.

[14] B. D. Loyd, "The effects of reality therapy/choice theory principles on high school students' perception of needs satisfaction and behavioral change." *International Journal of Reality Therapy*, vol. 25, no. 1, 2005.

[15] P. A. Robey, "Reality therapy and choice theory: An interview with robert wubbolding," *The Family Journal*, vol. 19, no. 2, pp. 231–237, 2011.

[16] O. Massah, F. Farmani, R. Karimi, H. Karami, F. Hoseini, and A. Farhoudian, "Group reality therapy in addicts rehabilitation process to reduce depression, anxiety and stress," *Iranian Rehabilitation Journal*, vol. 13, no. 1, pp. 50–44, 2015.

[17] W. Glasser, *Warning: Psychiatry can be hazardous to your mental health.* HarperCollins Publishers, 2003.

[18] A. Al-Kandari, S. R. Melkote, and A. Sharif, "Needs and motives of instagram users that predict self-disclosure use: A case study of young adults in kuwait," *Journal of Creative Communications*, vol. 11, no. 2, pp. 85–101, 2016.

[19] H. S. Hwang and J. Cho, "Why instagram? intention to continue using instagram among korean college students," *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. Torralba, "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 conference 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 Linguistics*, vol. 5, pp. 135–146, 2017.

[25] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1081–1094, 2018.

[26] Y. Kim and J. H. Kim, "Using computer vision techniques 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 systematic 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] 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.

[34] J. E. Hunter and F. L. Schmidt, *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

[35] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[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.

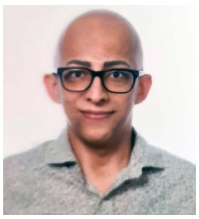
[44] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients." *Psychological methods*, vol. 1, no. 1, p. 30, 1996.

[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.

Bita 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 Autonoma de Barcelona. He received his Ph.D. in 2005 and was awarded for the best thesis in the Computer Science.