Multimodal Analysis and Prediction of Latent User Dimensions

Laura Wendlandt[1], Rada Mihalcea[1], Ryan L. Boyd[2], and James W. Pennebaker[2]

[1] University of Michigan, Ann Arbor, MI, USA,
wenlaura@umich.edu, mihalcea@umich.edu
[2] University of Texas at Austin, Austin, TX, USA,
ryanboyd@utexas.edu, pennebaker@utexas.edu

Abstract. Humans upload over 1.8 billion digital images to the internet each day, yet the relationship between the images that a person shares with others and his/her psychological characteristics remains poorly understood. In the current research, we analyze the relationship between images, captions, and the latent demographic/psychological dimensions of personality and gender. We consider a wide range of automatically extracted visual and textual features of images/captions that are shared by a large sample of individuals ($N \approx 1,350$). Using correlational methods, we identify several visual and textual properties that show strong relationships with individual differences between participants. Additionally, we explore the task of predicting user attributes using a multimodal approach that simultaneously leverages images and their captions. Results from these experiments suggest that images alone have significant predictive power and, additionally, multimodal methods outperform both visual features and textual features in isolation when attempting to predict individual differences.

Keywords: Analysis of latent user dimensions, Multimodal prediction, Joint language/vision models

1 Introduction

Personalized image data has become widespread: over 1.8 billion digital images are added to the internet each day [29]. Despite this tremendous quantity of visual data, the relationship between the images that a person shares online and his/her demographic and psychological characteristics remains poorly understood. One of the most appealing promises of this data is that it can be used to gain a deeper understanding into the thoughts and behaviors of people.

Specifically, in this work, we examine the relationship between images, their captions, and the latent user dimensions of personality and gender to address several basic questions. First, from a correlational perspective, how do image and caption attributes relate to the individual traits of personality and gender? We extract an extensive set of visual and textual features and use correlational techniques to uncover new, interpretable psychological insight into the ways that image attributes (such as objects, scenes, and faces) as well as language features (such as words and semantic categories) relate to personality and gender. Second, do image attributes have predictive power for

these traits? We demonstrate that visual features alone have significant predictive power for latent user dimensions. While previous work has extensively explored the connection between textual features and user traits, we are among the first to show that images can also be used to predict these traits. Finally, how can we combine visual and textual features in a multimodal approach to achieve better predictive results? We develop multimodal models that outperform both visual features and textual features in isolation when attempting to predict individual differences. We also show that these models are effective on a relatively small corpus of images and text, in contrast to other published multimodal approaches for tasks such as captioning, which rely on very large visual and textual corpora.

2 Related Work

When studying individuals, we are often trying to get a general sense of who they are as a person. These types of evaluations fall under the broader umbrella of *individual differences*, a large area of research that tries to understand the various ways in which people are psychologically different from one another, yet relatively consistent over time [2]. A large amount of research in the past decade has been dedicated to the assessment and estimation of individual characteristics as a function of various behavioral traces. In our case, these traces are images and captions collected from undergraduate students.

Personality Prediction. Much of the work in individual differences research focuses on the topic of *personality*. Generally speaking, "personality" is a term used in psychology to refer to constellations of feelings, behaviors, and cognitions that co-occur within an individual and are relatively stable across time and contexts. Personality is most often conceived within the Big 5 personality framework, and these five dimensions of personality are predictive of important behavioral outcomes such as marital satisfaction [16] and even health [36].

From a computational perspective, the problem of predicting personality has primarily been approached using Natural Language Processing (NLP) methods. While the textual component of our work focuses on short image captions, most previous research used longer bodies of text such as essays or social media updates [33]. N-grams, as well as psychologically-derived linguistic features such as those provided by LIWC, have been shown to have significant predictive power for personality [25, 34].

In addition to textual inference, there has been a recent movement towards incorporating images into the study of individual differences. Similar to our work, Segalin et al. have found that both traditional computer vision attributes and convolutional neural networks can be used to infer personality [38, 39]. Liu et al. have also discussed the possibility of inferring personality from social media profile pictures [22]. However, unlike our work, these studies do not make use of higher-level image features (e.g. scenes, objects), and they do not consider any image captions or any interaction between visual and textual modalities.

Gender Prediction. Contemporary research on individual differences extends well beyond personality evaluations to include variables such as gender, age, life experiences, and so on – facets that differ between individuals but are not necessarily caused by in-

ternal psychological processes. In addition to personality, we also consider gender in this work.

As with personality, predicting gender has primarily been approached using NLP techniques [18, 21, 31]. Relevant to the current work, however, is recent work by You et al. [43], who have explored the task of predicting gender given a user's selected images on Pinterest, an online social networking site.

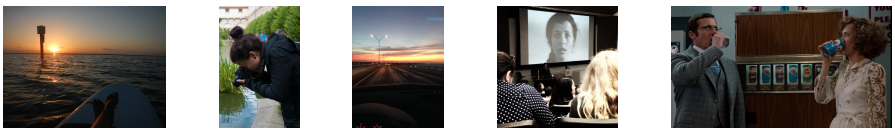
Inference from Multiple Modalities. Our work also relates to the recent body of research on the joint use of language and vision. Our multimodal approach is particularly related to automatic image annotation, the task of extracting semantically meaningful keywords from images [44]. Other related multimodal approaches can be found in the fields of image captioning [15] and joint text-image embeddings [3]. Some of these approaches rely on very large corpora. For example, Johnson et al. train an image captioning algorithm using Visual Genome, a dataset with greater than 94,000 images [14].

3 Dataset

We use a dataset collected at the University of Texas at Austin in the context of a Fall 2015 online undergraduate introductory psychology class.[3] The dataset includes free response data and responses to standard surveys collected from 1,353 students ages 16 to 46 (average 18.8 $\pm$ 2.10). The ethnicity distribution is 40.3% Anglo-Saxon/White, 27.1% Hispanic/Latino, 22.3% Asian/Asian American, 5.5% African American/Black, and 4.8% Other/Undefined.

Three elements of this dataset are of particular interest to our research:

Free Response Image Data. Each student was asked to submit and caption five images that expressed who he/she is as a person. As Fig. 1 illustrates, students submitted a wide range of images, from memes to family photos to landscapes. Some students chose to submit fewer than five images.



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

Fig. 1: Five images from the dataset submitted by a single student (with student faces blurred out for privacy). The accompanying captions are: (a) I'd rather be on the water. (b) The littlest things are always so pretty (and harder to capture). (c) I crossed this bridge almost every day for 18 years and never got tired of it. (d) The real me is right behind you. (e) Gotta find something to do when I have nothing to say.

[3] This data was collected under IRB approval at UT Austin.

Big 5 Personality Ratings. Each student completed the BFI-44 personality inventory, which is used to score individuals along each of the Big 5 personality dimensions using a 1-to-5 scale [13]. The Big 5 personality dimensions include [28]: *Openness* (example adjectives: artistic, curious, imaginative, insightful, original, wide interests); *Conscientiousness* (efficient, organized, planful, reliable, responsible, thorough); *Extraversion* (active, assertive, energetic, enthusiastic, outgoing, talkative); *Agreeableness* (appreciative, forgiving, generous, kind, sympathetic, trusting); and *Neuroticism* (anxious, self-pitying, tense, touchy, unstable, worrying).

Gender. Finally, demographic data is also associated with each student, including gender, which we use in our work. The gender distribution is 61.6% female, 37.8% male, and 0.5% undefined. Gender-unspecified students are omitted from our analyses.

Computing Correlations. An important contribution of our work is gathering new insights into textual and image attributes that correlate with personality and gender. Each of the personality dimensions is continuous, therefore, a version of the Pearson correlation coefficient is used to calculate correlations between personality and visual and textual features. Because there are, in some cases, thousands of image or text features, we must account for inferential issues associated with multiple testing (e.g., inflated error rates); we address such issues using a multivariate permutation test [42].

This approach is done by first calculating the Pearson product-moment correlation coefficient r for two variables. Then, for a high number of iterations (in our case, 10,000), the two variables are randomly shuffled and the Pearson coefficient is recalculated each time. At the end of the shuffling, a two-tailed p-test is conducted, comparing the true correlation r with the values of r attained from randomly shuffling the data. The original result is considered to be legitimate only when the original Pearson's r is found to be statistically significant in comparison to all of the random coefficients. As discussed in [42], for small sample sizes, this multivariate permutation test has more statistical power than the common Bonferroni correction.

Unlike personality, gender is a categorical variable. Thus, Welch's t-tests are used to look for significant relationships between gender and image and text features. These relationships are measured using effect size (Cohen's d), which measures how many standard deviations the two groups differ by, and is calculated by dividing the mean difference by the pooled standard deviation. In using Welch's t-tests, we make the assumption that within each gender, image and text features follow a normal distribution.

4 Analysing Images

In order to explore the relationship between images and psychological attributes, we want to extract meaningful and interpretable image features that have some connection to the user.

4.1 Raw Visual Features

We begin by describing low-level raw visual features of an image.

Colors. Past research has shown that colors are associated with abstract concepts [35]. For instance, red is associated with excitement, yellow with cheerfulness, and blue with

comfort, wealth, and trust. Furthermore, research has shown that men and women respond to color differently. In particular, one study found that men are more tolerant of gray, white, and black than are women [17].

To characterize the distribution of colors in an image, we classify each pixel as one of eleven named colors using the method presented by Van De Weijer et al. [41]. This method trains a Probabilistic Latent Semantic Analysis model over retrieved Google images, using the model to assign color names to individual pixels. For our experiments, we use Van De Weijer et al.'s pre-trained model. The percentage of each color across an image is used as a feature.

Brightness and Saturation. Images are often characterized in terms of their brightness and saturation. Here, we use the HSV color space, where brightness is defined as the relative lightness or darkness of a particular color, from black (no brightness) to light, vivid color (full brightness). Saturation captures the relationship between the hue of a color and its brightness and ranges from white (no saturation) to pure color (full saturation). We calculate the mean and the standard deviation for both the brightness and the saturation.

Previous work has also used brightness and saturation to calculate metrics measuring pleasure, arousal, and dominance, as expressed in the following formulas: $Pleasure = 0.69y + 0.22s$; $Arousal = -0.31y + 0.60s$; $Dominance = -0.76y + 0.32s$, where y is the average brightness of an image and s is its average saturation [40].[4]

Texture. The texture of an image provides information about the patterns of colors or intensities in the image. Following [23], we use Grey Level Co-occurrence Matrices (GLCMs) to calculate four texture metrics: contrast, correlation, energy, and homogeneity.

Static and Dynamic Lines. Previous work has shown that the orientation of a line can have various emotional effects on the viewer [24]. For example, diagonal lines are associated with movement and a lack of equilibrium. To capture some of these effects, we measure the percentage of static lines with respect to all of the lines in the image.[5] Static lines are defined as lines that are within $\pi/12$ radians of being vertical or horizontal.

Circles. The presence of circles and other curves in images has been found to be associated with emotions such as anger and sadness [35]. Following the example of [35], we calculate the number of circles in an image.[6]

Correlations. Once the entire set of raw features is extracted from the images, correlations between raw features and personality/demographic features are calculated. Table 1 presents significant correlations between visual features and personality traits. One correlation to note is a positive relationship between the number of circles in an image and extraversion. This is likely because the circle detection algorithm often counts

[4] For prediction experiments, we use a slightly different version of dominance ($Dominance = 0.76y + 0.32s$), as formulated in [24].

[5] We use the OpenCV probabilistic Hough transform function with an accumulator threshold of 50, a minimum line length of 50, and a maximum line gap of 10.

[6] We use the OpenCV Hough circles function, with a minimum distance of 8 and method-specific parameters set to 170 and 45.

faces as circles, and faces have a natural connection with the social facets of extraversion. Our results also validate the findings of Valdez and Mehrabian, who suggest that pleasure, arousal, and dominance have emotional connections [40]. Here we show that these metrics also have connections to personality. While these correlations are weak, they are statistically significant.

Table 1: Significant correlations between image attributes and Big 5 personality traits. These correlations are corrected using a multivariate permutation test, as described in Section 3. Only scenes and basic WordNet domains that have one of the top five highest correlations or one of the top five lowest correlations are shown.

	Big 5 Personality Dimensions				
Image Attributes	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Raw Visual Features					
Black	-	-	-	-0.06	-
Blue	-	-	0.06	-	-0.07
Grey	0.06	-	-0.11	-	-
Orange	-	-	0.07	-	-
Purple	-	-	0.06	-	-
Red	-0.06	-	-	-	-
Brightness Std. Dev.	-	-	0.07	-	-
Saturation Mean	-	-	0.07	-	-0.06
Saturation Std. Dev.	-0.06	-	0.06	-	-0.06
Pleasure	-	-	0.07	-	-0.05
Arousal	-	-	-	-	0.06
Dominance	-	-	0.08	-	-0.05
Homogeneity	-	0.05	-	-	-
Static Lines %	-	-	-0.07	-	-
Num. of Circles	-	-	0.10	-	-0.06
Scenes					
Ballroom	-	-	0.12	0.06	-0.06
Bookstore	-	-0.06	-0.11	-	0.08
Canyon	0.11	-	-	-	-
Home Office	-	-	-0.12	-	-
Mansion	-	-	-	0.10	-
Martial Arts Gym	-0.09	-	0.06	-	-
Pantry	-	-	-0.11	-0.10	-
Playground	-	-	0.07	0.09	-0.06
River	-	0.09	0.07	0.07	-
Shower	-	-	-0.09	-	-
Faces	-0.07	0.08	0.17	0.11	-
WordNet Supersenses					
Animal	-	-	0.06	-	-
Person	-	-	-	-	-0.06
Basic WordNet Domains					
History	-	-	0.06	-	-
Play	-0.10	-	-	-	-
Sport	-0.10	-	-	-	-
Home	-	-0.06	-0.09	-	-
Biology	-	-	0.07	-	-
Physics	-	-0.08	-	-0.09	-
Anthropology	-	0.06	-	-	-
Industry	-	-	-0.08	-	-
Fashion	-0.07	0.06	0.11	0.05	-

Table 2 shows effect sizes for features significantly different between men and women. As suggested by previous research, men are more likely to use the color black [17]; other correlations appear to confirm stereotypes, e.g., a stronger preference by women for pink and purple.

Table 2: Image and text features where there is a significant difference ($p < 0.05$) between male and female images. Only scenes, basic WordNet domains, and unigrams with the highest ten effect sizes (by magnitude) are shown. All text features except for the word count itself are normalized by the word count. Positive effect sizes indicate that women prefer the feature, while negative effect sizes indicate that men prefer the feature.

Image Attributes	Effect Size
Raw Visual Features	
Pink	0.455
Static Lines %	-0.360
Black	-0.325
Brightness Mean	0.266
Saturation Std. Dev.	-0.176
Purple	0.167
Brown	0.166
Homogeneity	0.118
Red	0.111
Faces	0.160
Scenes	
Beauty Salon	0.347
Ice Cream Parlor	0.340
Office	-0.290
Slum	0.286
Football Stadium	-0.267
Basement	-0.235
Herb Garden	0.224
Gas Station	-0.222
Music Studio	-0.222
Baseball Stadium	-0.222
WordNet Supersenses	
Artifact	-0.213
Person	-0.173
Food	0.107
Basic WordNet Domains	
Play	-0.236
Sport	-0.235
Transport	-0.186
Military	-0.182
Animals	-0.155
History	-0.153
Art	-0.142
Food	0.136
Plants	0.120
Tourism	-0.118

Text Attributes	Effect Size
Stylistic Features	
Num. of Words	0.174
Readability - GFI	-0.161
Readability - SMOG	-0.146
Readability - FRE	-0.136
Unigrams	
Boyfriend	0.361
Girlfriend	-0.360
Was	0.287
Play	-0.285
She	0.264
Them	0.262
Sport	-0.254
Sister	0.244
Game	-0.242
Enjoy	-0.236
LIWC Categories	
Prepositions	-0.198
Past Focus	0.176
Sports	-0.173
Work	-0.167
Period	-0.157
Other References	0.145
Quote	-0.133
Other	0.123
1st Person Plural Personal Pronouns	0.123
MRC Categories	
Kucera-Francis Written Freq.	-0.139
Kucera-Francis Num. of Samples	-0.134

4.2 Scenes

Previous research has linked personal spaces (such as bedrooms and offices) with various personality attributes, indicating that how a person composes his/her space provides clues about his/her psychology, particularly through self-presentation and related social processes [11].

In order to identify the scene of an image, we use Places-CNN [45], a convolutional neural network (CNN) trained on approximately 2.5 million images and able to classify an image into 205 scene categories. To illustrate, Fig. 2 shows sample images. For each image, we use the softmax probability distribution over all scenes as features.

(a) Coffee Shop (0.53), Ice Cream Parlor (0.24)

(b) Parking Lot (0.57), Sky (0.26)

Fig. 2: Top scene classifications for two images, along with their probabilities.

Correlations. Scenes strongly correlated with personality traits are shown in Table 1. The strongest positive correlation is between extraversion and ballrooms, and the strongest negative correlation is between extraversion and home offices. Findings such as these are conceptually sound, as individuals tend to engage in personality-congruent behaviors. In other words, individuals scoring high on extraversion are expected to feel that inherently social locations, such as ballrooms, are more relevant to the self than locations indicative of social isolation, such as home offices.

We also measure the relationship between scenes and gender. Table 2 shows scenes that are associated with either males or females. Men are more commonly characterized by sports-related scenes, such as football and baseball stadiums, whereas women are more likely to have photos from ice cream and beauty parlors. As illustrated in Fig. 2, the scene detection algorithm tends to conflate coffee shops and ice cream parlors, so this observed preference for ice cream parlors could be partially attributed to a preference for coffee shops.

4.3 Faces

Most aspects of a person's personality are expressed through social behaviors, and the number of faces in an image can capture some of this behavior. We use the work by Mathias et al. to detect faces [27]. Specifically, we use their pretrained *HeadHunter* model, an advanced evolution of the Viola-Jones detector.

Correlations. Significant correlations between faces and personality traits are shown in Table 1. Of particular note is the strong positive correlation between the number of faces and extraversion, which is intuitive because extraverts are often thought of as enjoying social activities. With respect to gender, Table 2 shows that women tend to have more faces in their images than men.

4.4 Objects

Previous research has indicated that people can successfully predict other's personality traits by observing their possessions [10]. This indicates that object detection has the ability to capture certain psychological insight.

To detect multiple objects per image, we break each image into multiple regions, apply object detection to each region, and post-process each detection to create interpretable features. To identify image regions, we use *Edge Boxes* [46] to detect a maximum of 2,000 regions;[7] to avoid falsely detecting objects, regions that are less than 3% of the total image area are discarded.

Objects are detected by sending each region through CaffeNet, a version of AlexNet [12, 19]. CaffeNet assumes that each region contains one object and outputs a softmax probability over 1,000 ImageNet objects [37]. The final score for an object in a region is the *Edge Boxes* score for the region multiplied by CaffeNet's softmax probability. We remove any objects with a score below a certain threshold, where the threshold is optimized on the PASCAL VOC image set [7]. For each object type, the scores of all of the detected objects in a particular image are added up, creating a 1,000-dimensional feature vector.

Because of the small size of our dataset and the large number of ImageNet objects, this feature vector is somewhat sparse and hard to interpret. To increase interpretability for correlational analysis, we consider two coarser-grained systems of classification: WordNet supersenses and WordNet domains. WordNet [8] is a large hierarchical database of English concepts (or synsets), and each ImageNet object is directly associated with a WordNet concept. Supersenses are broad semantic classes labeled by lexicographers (e.g., communication, object, animal) [5]. WordNet domains [1] is a complementary synset labeling. It groups WordNet synsets into various domains, such as medicine, astronomy, and history. The domain structure is hierarchical, but here we consider only basic WordNet domains, which are domains that are broad enough to be easily interpretable (e.g., history, chemistry, fashion). An object is allowed to fall into more than one domain.

Correlations. WordNet supersenses and domains correlate significantly with multiple personality traits, as shown in Table 1. Table 2 shows object classes that are different for males and females. These object classes connect back to scenes associated with men and women. For example, men are more likely to have sports objects in their images, reflected in the fact that men are more likely to include scenes of sports stadiums.

[7] We use the *Edge Boxes* parameters $\alpha = 0.65$ and $\beta = 0.55$.

4.5 Captions

When available, captions can be considered another way of representing image content via a textual description of the salient objects, people, or scenes in the image. Importantly, the captions have been contributed by the same people who contributed the images, and they represent the views that the image "owners" have about their content.

Stylistic Features. To capture writing style, we consider surface-level stylistic features, such as the number of words and the number of words longer than six characters. We also use the Stanford Named Entity Recognition system to extract the number of references to people, locations, and organizations [9]. Finally, we look at readability and specificity metrics. For readability, we consider a variety of metrics: Flesch Reading Ease (FRE), Automated Readability Index (ARI), Flesch-Kincaid Grade Level (FK), Coleman-Liau Index (CLI), Gunning Fog Index (GFI), and SMOG score (SMOG). For specificity, we use Speciteller [20].

N-grams. In addition to style, we want to capture the content of each caption. We do this by considering unigrams, bigrams, and trigrams. Each caption is tokenized (split into tokens on punctuation other than periods) and stemmed using the Lancaster Stemmer [4]. Only n-grams that occur more than five times are considered. N-grams that occur less than this are replaced by an out-of-vocabulary (OOV) symbol. We also consider part-of-speech (POS) unigrams, bigrams, and trigrams, tokenized using the Penn Treebank tagset [26].

LIWC Features. Linguistic Inquiry and Word Count (LIWC) is a word-based text analysis program [34]. It focuses on broad categories such as language composition, as well as emotional, cognitive, and social processes. We analyze each piece of text using LIWC in order to capture psychological dimensions of writing. For each of the 86 LIWC categories, we calculate a feature that reflects the percentage of caption words belonging to that category.

MRC Features. The MRC Psycholinguistic Database contains statistics about word use [6]. MRC features are calculated by averaging the values of all of the words in the caption. In our correlational analysis, certain MRC features emerge as particularly relevant. These include word frequency counts, which capture how common a word is in standard English usage, as well as measures for meaningfulness, imagery, and length (e.g., number of letters, phenomes, and syllables). These features provide a complementary perspective to the LIWC features.

Word Embeddings. For our prediction tasks, we also consider each word's embedding. *Word2vec* (*w2v*) is a method for creating a multidimensional embedding for a particular word [30]. Google provides pre-trained word embeddings on approximately 100 billion words of the Google News dataset.[8] For each caption in our dataset, we average together all of the word embeddings to produce a single feature vector of length 300. We use the Google embeddings for this, discarding words that are not present in the pre-trained embeddings.

Correlations. For correlational analysis, we normalize all text features by word count. Table 3 shows correlations between language features and personality. Interestingly,

[8] Available at https://code.google.com/archive/p/word2vec/.

there are very few strong correlations for extraversion. This is complementary to what we see with images, where there are many strong correlations for extraversion, suggesting that we are gleaning different aspects of personality from both images and text. Many of these textual correlations have been discussed in previous literature (e.g. [34, 25]), and our work confirms previous results.

Table 2 shows language features that are different between men and women. Things to note here are that women tend to write longer captions and men again exhibit a preference for talking about sports.

5 Multimodal Prediction

The task of prediction can provide valuable insights into the relationship between images, captions, and user dimensions. Here, we consider six coarse-grained classification tasks, one for each personality trait and one for gender. For each prediction, we divide the data into high and low segments. The high segment includes any person who has a score greater than half a standard deviation above the mean, while the low segment includes any person who has a score lower than half a standard deviation below the mean. All other data points are discarded. This binary division of personality traits results in mostly balanced data, with the high segment for each trait containing 47.7-51.8% of the data points. For gender, 61.6% of the data is female. In doing these coarse-grained classification tasks, we follow previous work [25, 32], which suggested that classification serves as a useful approximation to continuous rating.

We use a random forest with 500 trees and 10-fold cross validation across individuals in the dataset. Table 4 shows the classification results. As a baseline, we include a model that always predicts the most common training class. In addition to the random forest model, we also considered other approaches to this problem, primarily neural network-based. These approaches were not successful, partially because of the small size of our dataset, though they suggest some interesting future avenues to explore.

To enable direct comparison to previously published results, we use our data to retrain the personality prediction models from Mairesse et al. [25]; the re-trained classifier with the highest accuracies on our data, SMO, is shown in Table 5. We also include the relative error rate reduction between this model and our best multimodal model.

Single Modality Methods. To understand the predictive power of images and captions individually, we consider a series of predictions using feature sets derived from either only visual data or only textual data. These feature sets are the same features that we described in Section 4.

As shown in Table 4, image features in isolation are able to significantly classify both extraversion and gender. Text features are also able to significantly classify these traits, with slightly less accuracy than image features. Text features have additional predictive power for openness.

Multimodal Methods. We experiment with several methods of combining visual and textual data. First, we concatenate both the image and text feature vectors (excluding *w2v* embeddings). These results are shown in Table 4 under the *All* row in the *Image and Caption Attributes* section.

Table 3: Significant correlations between language attributes and Big 5 personality traits. All features except for the word count itself are normalized by the word count. Only unigrams, LIWC categories, and MRC categories that have one of the top five highest correlations or one of the top five lowest correlations are shown.

	Big 5 Personality Dimensions				
Language Attributes	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Stylistic Features					
Num. of Words	0.14	-	-	-	0.07
Words Longer than Six Characters	-	0.09	0.06	-	-
Num. of Locations	-	0.07	-	-	-0.07
Readability - FRE	-0.13	-	-	-	-
Readability - ARI	-	0.06	-	-	-
Readability - GFI	-0.14	-	-	-	-
Readability - SMOG	-0.13	-	-	-	-0.06
Specificity	-	0.08	-	-	-0.06
Unigrams					
Decide	-	-0.12	-	-	-
Different	0.11	-	-	-	-
In	0.11	-	-	-	-
It	0.11	-	-	-	-
King	-	0.06	-	-0.15	-
Level	-	-	-0.12	-	-
My	-0.14	0.07	-	-	-
Photoshop	0.10	-	-	-	-
Sport	-0.14	-	-	-	-
Write	0.10	-	-	-	-
LIWC Categories					
Achievement	-	0.08	-	-	-
All Punctuation	-	0.08	-	-	-0.07
Discrepancies	-	-	0.10	-	-0.07
1st Person Singular Personal Pronouns	-0.10	-	-	-	-
Inclusive	-	-	-	0.08	-
Occupation	-	0.08	0.06	-	-
Other References	-0.10	-	-	-	-0.06
1st Person Personal Pronouns	-0.10	-	-	-	-
Sports	-0.11	0.07	-	-	-
Unique	-	0.07	-	0.08	-0.09
MRC Categories					
Imagery	-0.07	0.06	-	0.06	-0.07
Kucera-Francis Num. of Categories	-0.07	0.06	-	0.07	-0.09
Kucera-Francis Num. of Samples	-0.08	-	-	-	-0.07
Mean Pavio Meaningfulness	-0.08	-	-	-	-0.07
Num. of Letters in Word	-	0.08	-	0.07	-0.08
Num. of Phonemes in Word	-	0.08	-	0.07	-0.08
Num. of Syllables in Word	-	0.08	-	0.08	-0.08

To provide a more nuanced combination of features, we introduce the idea of image-enhanced unigrams (IEUs). This is a bag-of-words representation of both an image and its corresponding caption. It includes all of the caption unigrams, as well as unigrams derived from the image. We consider two methods, macro and micro, for generating image unigrams. For the macro method, we examine each individual image. If a color covers more than one-third of the image, the name of the color is added to the bag-of-words. The scene with the highest probability and any objects detected in the image

Table 4: Classification accuracy percentages. O, C, E, A, and N stand for openness, conscientiousness, extraversion, agreeableness, and neuroticism, respectively. * indicates significance with respect to the baseline ($p < 0.05$). Only image features that produce significant results and text features that score highest in one of the categories are shown.

	Predicted Attributes					
Feature Set Used	O	C	E	A	N	Gender
Baseline: Most Common Class	51.4	52.0	49.2	51.8	52.3	59.8
Image Attributes Only						
Object	**55.6**	51.7	57.2*	51.3	51.7	64.7*
Scene	55.5	53.8	59.8*	55.0	55.2	66.8*
Face	50.7	51.2	58.5*	54.1	51.1	59.7
All	54.8	**54.3**	**59.9***	**55.3**	**55.3**	**68.6***
Caption Attributes Only						
Unigrams	60.2*	53.1	54.3	54.2	53.4	**67.6***
Bigrams	58.0*	**53.2**	**57.6***	53.4	**57.3**	65.1*
LIWC	59.6*	**53.2**	54.1	53.4	54.2	64.9*
All (except pre-trained *w2v*)	61.2*	52.2	53.3	54.6	55.2	65.1*
Pre-trained *w2v* (caption only)	**61.8***	51.4	55.4	**55.4**	56.5	67.1*
All + Pre-trained *w2v* (caption only)	61.2*	52.3	55.5	53.0	56.1	65.6*
Image and Caption Attributes						
All	60.5*	55.1	57.9*	55.3	56.8	67.1*
Macro IEU	58.5*	56.6	58.5*	54.2	54.7	**71.0***
Micro IEU	58.7*	54.4	58.9*	54.0	52.7	**71.0***
All + Macro IEU	60.0*	**57.1**	58.3*	54.2	56.9	68.1*
All + Micro IEU	59.1*	55.6	60.3*	54.8	**58.3***	69.1*
Pre-trained *w2v* (w/ Micro IEU)	**61.4***	54.8	59.6*	56.4*	56.5	68.6*
Pre-trained *w2v* (w/ Macro IEU)	61.0*	55.6	**60.5***	**57.0***	56.6	69.0*
All + Pre-trained *w2v* (w/ Micro IEU)	59.5*	54.8	59.1*	55.3	55.3	70.1*
All + Pre-trained *w2v* (w/ Macro IEU)	**61.4***	54.7	59.4*	55.2	56.5	70.8*

Table 5: Comparison between our best classification model and the best model (SMO) from Mairesse et al. * indicates significance with respect to the baseline ($p < 0.05$).

	Predicted Attributes					
Feature Set Used	O	C	E	A	N	Gender
Baseline: Most Common Class	51.4	52.0	49.2	51.8	52.3	59.8
Mairesse et al.: SMO	59.1*	51.3	53.3	54.4	54.7	63.0
Our model: Pre-trained *w2v* (w/ Macro IEU)	61.0*	55.6	60.5*	57.0*	56.6	69.0*
Relative error rate reduction (our model vs. Mairesse et al.)	4.6%	8.8%	15.4%	5.7%	4.2%	16.2%

are also added. The unigrams from each individual image are then combined with the caption unigrams to form the set of macro IEUs. To generate micro IEUs, instead of considering individual images, we consider aggregated image characteristics. First, for each student, we average his/her image feature vectors into a single vector, and then we extract image unigrams from this combined vector. For example, if the average percentage of a particular color across all images is greater than 33%, the name of that color is added to the bag-of-words. These unigrams are mixed with the caption unigrams to form the set of micro IEUs.

We use IEUs in several different ways for prediction. First, we consider them both in isolation and concatenated with all of the previous visual and textual features (excluding *w2v*). We also explore using the pre-trained *w2v* model to represent the IEUs and produce richer embeddings. Instead of only averaging together the embeddings of each caption unigram, we average together the embeddings of each IEU. Finally, we consider these enriched embeddings concatenated with all of the previous features.

A significant advantage of these multimodal approaches is that they can be used with relatively small corpora of images and text. Large background corpora are used for training (e.g., for training the scene CNN), but these models have already been trained and released. Our approaches work when there is only a small amount of training data, as is often the case when ground truth labels are expensive to obtain. This is demonstrated on our dataset, which consists of short captions and a relatively limited set of images.

The results obtained with the multimodal methods are shown in the bottom part of Table 4. As seen in the table, when compared to the methods that rely on individual modalities, these multimodal models outperform image features in all six categories and text features in all but one category. The methods using IEUs achieve the best results and are able to significantly classify both neuroticism and agreeableness, something that neither visual features nor textual features are able to do in isolation.

As shown in Table 5, our multimodal approaches also outperform the method from Mairesse et al., achieving relative error rate reductions between 4% and 16%.

6 Conclusion

This paper explores the connection between images, captions, and the latent user dimensions of personality and gender. While there is a large body of previous work that has considered the use of text as a way to analyse and predict user traits, there is very little work on the use of images for this task. The paper makes several contributions. First, using a new dataset of captioned images associated with user attributes, we extract a large set of visual and textual features and identify significant correlations between these features and the user traits of personality and gender. Second, we demonstrate the effectiveness of image features in predicting user attributes; we believe this result can have applications in many areas of the web where textual data is limited. Finally, we show that a multimodal predictive approach outperforms purely visual and textual methods. Our multimodal methods are also effective on relatively small corpora.

Acknowledgments This material is based in part upon work supported by the National Science Foundation (NSF #1344257), the John Templeton Foundation (#48503), and the Michigan Institute for Data Science (MIDAS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, the John Templeton Foundation, or MIDAS. We would like to thank Chris Pittman for his aid with the data collection, Shibamouli Lahiri for the readability code, and Steven R. Wilson for the implementation of Mairesse et al.

References

1. Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising the wordnet domains hierarchy: semantics, coverage and balancing. In: Proceedings of the Workshop on Multilingual Linguistic Ressources. pp. 101–108. Association for Computational Linguistics (2004)
2. Boyd, R.L.: Psychological text analysis in the digital humanities. In: Hai-Jew, S. (ed.) Data Analytics in the Digital Humanities. Springer Science. In press., New York City, NY (2017)
3. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. Journal of Artificial Intelligence Research 49(1-47) (2014)
4. Chris, D.P.: Another stemmer. In: ACM SIGIR Forum. vol. 24, pp. 56–61 (1990)
5. Ciaramita, M., Johnson, M.: Supersense tagging of unknown nouns in WordNet. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 168–175. Association for Computational Linguistics (2003)
6. Coltheart, M.: The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology 33(4), 497–505 (1981)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
8. Fellbaum, C.: WordNet. Wiley Online Library (1998)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370 (2005)
10. Gosling, S.D., Craik, K.H., Martin, N.R., Pryor, M.R.: Material attributes of personal living spaces. Home Cultures 2(1), 51–87 (2005)
11. Gosling, S.D., Ko, S.J., Mannarelli, T., Morris, M.E.: A room with a cue: Personality judgments based on offices and bedrooms. Journal of Personality and Social Psychology 82(3), 379 (2002)
12. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. pp. 675–678 (2014)
13. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of Personality: Theory and Research 2(1999), 102–138 (1999)
14. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: Fully convolutional localization networks for dense captioning. arXiv preprint arXiv:1511.07571 (2015)
15. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137 (2015)
16. Kelly, E.L., Conley, J.J.: Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. Journal of Personality and Social Psychology 52(1), 27 (1987)
17. Khouw, N.: The meaning of color for gender. Colors Matters–Research (2002)
18. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
20. Li, J.J., Nenkova, A.: Fast and accurate prediction of sentence specificity. In: AAAI. pp. 2281–2287 (2015)
21. Liu, H., Mihalcea, R.: Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In: International Conference on Weblogs and Social Media (2007)

22. Liu, L., Preotiuc-Pietro, D., Samani, Z.R., Moghaddam, M.E., Ungar, L.: Analyzing personality through social media profile picture choice. In: Tenth International AAAI Conference on Web and Social Media (2016)
23. Lovato, P., Bicego, M., Segalin, C., Perina, A., Sebe, N., Cristani, M.: Faved! Biometrics: Tell me which image you like and I'll tell you who you are. IEEE Transactions on Information Forensics and Security 9(3), 364–374 (2014)
24. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM International Conference on Multimedia. pp. 83–92. ACM (2010)
25. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30, 457–500 (2007)
26. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics 19(2), 313–330 (1993)
27. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision. pp. 720–735. Springer (2014)
28. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. Journal of Personality 60(2), 175–215 (1992)
29. Meeker, M.: Internet trends 2014–Code conference (2014), retrieved May 28, 2014
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
31. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes 45(3), 211–236 (2008)
32. Oberlander, J., Nowson, S.: Whose thumb is it anyway?: Classifying author personality from weblog text. In: COLING/ACL. pp. 627–634 (2006)
33. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P.: Automatic personality assessment through social media language. Journal of Personality and Social Psychology (2014)
34. Pennebaker, J.W., King, L.A.: Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology 77(6), 1296 (1999)
35. Redi, M., Quercia, D., Graham, L., Gosling, S.: Like partying? Your face says it all. Predicting the ambiance of places with profile pictures. In: Ninth International AAAI Conference on Web and Social Media (2015)
36. Roberts, B., Kuncel, N., Shiner, R., Caspi, A., Goldberg, L.: The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 4(2), 313–345 (2007)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
38. Segalin, C., Cheng, D.S., Cristani, M.: Social profiling through image understanding: Personality inference using convolutional neural networks. Computer Vision and Image Understanding (2016)
39. Segalin, C., Perina, A., Cristani, M., Vinciarelli, A.: The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Transactions on Affective Computing (2016)
40. Valdez, P., Mehrabian, A.: Effects of color on emotions. Journal of Experimental Psychology: General 123(4), 394 (1994)
41. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing 18(7), 1512–1523 (2009)

42. Yoder, P.J., Blackford, J.U., Waller, N.G., Kim, G.: Enhancing power while controlling family-wise error: An illustration of the issues using electrocortical studies. Journal of Clinical and Experimental Neuropsychology 26(3), 320–331 (2004)
43. You, Q., Bhatia, S., Sun, T., Luo, J.: The eyes of the beholder: Gender prediction using images posted in online social networks. In: 2014 IEEE International Conference on Data Mining Workshop. pp. 1026–1030. IEEE (2014)
44. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition 45(1), 346–362 (2012)
45. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems. pp. 487–495 (2014)
46. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision. pp. 391–405. Springer (2014)