Received: 29 November 2023 | Accepted: 1 June 2024

DOI: 10.1111/bjop.12717

the british psychological society
promoting excellence in psychology

ARTICLE

Decoding the language of first impressions: Comparing models of first impressions of faces derived from free-text descriptions and trait ratings

Alex L. Jones[1] | **Victor Shiramizu**[2] | **Benedict C. Jones**[2]

[1]School of Psychology, Swansea University, Swansea, UK

[2]Department of Psychological Sciences & Health, University of Strathclyde, Glasgow, Scotland

Correspondence
Alex L. Jones, School of Psychology, Swansea University, Singleton Park, Swansea SA28PP, UK.
Email: alex.l.jones@swansea.ac.uk

Abstract

First impressions formed from facial appearance predict important social outcomes. Existing models of these impressions indicate they are underpinned by dimensions of Valence and Dominance, and are typically derived by applying data reduction methods to explicit ratings of faces for a range of traits. However, this approach is potentially problematic because the trait ratings may not fully capture the dimensions on which people spontaneously assess faces. Here, we used natural language processing to extract 'topics' directly from participants' free-text descriptions (i.e., their first impressions) of 2222 face images. Two topics emerged, reflecting first impressions related to positive emotional valence and warmth (Topic 1) and negative emotional valence and potential threat (Topic 2). Next, we investigated how these topics were related to Valence and Dominance components derived from explicit trait ratings. Collectively, these components explained only ~44% of the variance in the topics extracted from free-text descriptions and suggested that first impressions are underpinned by correlated valence dimensions that subsume the content of existing trait-rating-based models. Natural language offers a promising new avenue for understanding social cognition, and future work can examine the predictive utility of natural language and traditional data-driven models for impressions in varying social contexts.

KEYWORDS

computational modelling, methodology, person perception, perception, social cognition

BACKGROUND

First impressions based on facial appearance play an important role in social interaction (Todorov et al., 2015; Zebrowitz & Montepare, 2008). Impressions of faces on a wide range of traits (e.g., trustworthiness, competence and attractiveness) influence real-world social outcomes (Langlois et al., 2000; Rhodes, 2006; Todorov et al., 2015) and are formed rapidly and automatically (Borkenau et al., 2009; Eggleston et al., 2021; Olivola & Todorov, 2010; Willis & Todorov, 2006). For example, first impressions based on facial appearance predict election outcomes (Ballew & Todorov, 2007; Olivola & Todorov, 2010; Olivola et al., 2012), judicial decisions (Dumas & Testé, 2006; Olivola, Funk, & Todorov, 2014; Wilson & Rule, 2015), partner choice (South Palomares & Young, 2018), hiring decisions (Olivola, Eubanks, & Lovelace, 2014; Zebrowitz & Montepare, 2008) and many aspects of economic exchange (Gheorghiu et al., 2017; Jaeger et al., 2019; Menegatti et al., 2021; van't Wout & Sanfey, 2008). Thus, understanding the processes and mechanisms that underlie first impressions of faces can provide insight into an important driver of social interactions and outcomes (Olivola, Funk, & Todorov, 2014; Todorov et al., 2015; Zebrowitz & Montepare, 2008).

Research on first impressions formed from faces has investigated a diverse range of traits, such as trustworthiness (Jaeger et al., 2018; Stirrat & Perrett, 2010; Todorov, 2008; Todorov & Duchaine, 2008), dominance (Fruhen et al., 2015; Hester et al., 2021; Quist et al., 2011), attractiveness (Holzleitner et al., 2019; Jones & Jaeger, 2019), competence (Oh et al., 2019; Todorov et al., 2005), aggressiveness (Lefevre & Lewis, 2014), intelligence (Kleisner et al., 2014; Talamas et al., 2016) and sociability (Jaeger et al., 2022; Mehu et al., 2007). However, arguably the most important advances in the first-impressions literature have come from studies that used data-reduction methods to identify the core perceptual dimensions that underpin the wide range of often intercorrelated traits that are typically considered in the literature (Todorov et al., 2013). Moreover, Oosterhof and Todorov's (2008) 'valence–dominance' model, which was derived using this approach, has been particularly influential in the first-impressions literature.

The initial description of the valence–dominance model (Oosterhof & Todorov, 2008) first asked 55 participants to provide spontaneous descriptions for each of 66 faces. The researchers then sorted the 1134 descriptions produced at this initial step into 14 trait categories and, importantly for the current work, removed any descriptors that were unrelated to personality traits. They then assigned each of these categories a trait label, and had the 66 faces rated for each of these labels (plus the additional trait 'dominance') by a different group of participants. Principal Component Analysis (PCA) of these explicit trait ratings revealed two components that explained 81.6% of the variance in the explicit trait ratings. The first component, labelled Valence, was highly correlated with ratings of traits such as trustworthiness, emotional stability, responsibility, caringness and sociability, and was interpreted as reflecting impressions of other people's prosocial intentions. The second component, labelled Dominance, was highly correlated with ratings of traits such as dominance, aggressiveness and meanness, and was interpreted as reflecting impressions of other peoples' capacity to inflict harm on the perceiver. Although studies of explicit ratings of faces, bodies and voices in which stimuli were rated for the same (or conceptually similar) traits to those considered by Oosterhof and Todorov (2008) have also concluded that first impressions are underpinned by Valence and Dominance dimensions (Jones & Kramer, 2021; McAleer et al., 2014; Morrison et al., 2017; Shiramizu et al., 2022; Tzschaschel et al., 2022), some other studies that investigated different traits and/or used different data-reduction methods have observed additional dimensions (e.g., a youthfulness dimension) that also contributed to first impressions (Lin et al., 2021; Sutherland et al., 2013). Evidence for the valence–dominance model has also come from cross-cultural research, at least when the analytical methods employed were the same as those used by Oosterhof and Todorov (Jones et al., 2021).

A potentially important limitation of the studies described above is that, in each case, the dimensions underpinning first impressions were identified by analysing explicit trait ratings of stimuli (i.e., participants were explicitly instructed to rate stimuli for traits that were specified by the researchers). However, there is growing concern that such traits may not necessarily fully capture the richness of first impressions, as researchers decide what traits participants will rate the faces for (Mondloch et al., 2023; Satchell et al., 2022; Sutherland & Young, 2022). Indeed, some researchers have argued that it may be

more useful to analyse types of responses or data that are less constrained, and potentially influenced, by researcher expectations and assumptions (Jack et al., 2018; Satchell et al., 2022).

In light of the above, we first used natural language processing techniques (NLP) to analyse participants' free-text descriptions of 2222 face images (Study 1). These techniques derive statistical regularities from text data, extracting patterns of words or phrases that occur in documents (Ding et al., 2023; Liu et al., 2021) that represent the core 'topics' that might underpin participants' written descriptions (i.e., first impressions) of the individuals shown in the images. In Study 2, we investigated the extent to which the extracted topics could be predicted from valence and dominance components derived from the same explicit trait ratings considered in previous studies of the valence–dominance model. We carried out Study 2 to quantify the similarity between the topics extracted from free-text descriptions of faces in Study 1 and the valence and dominance components derived from explicit trait ratings.

STUDY ONE – TOPIC MODELLING OF FIRST IMPRESSIONS

To identify the core dimensions (or topics) underpinning free text descriptions (i.e., first impressions) of faces, we employed topic modelling with a large, open-access face image set (the 10k US Adult Faces Database, Bainbridge et al., 2013). This approach allows us to use topic modelling in an analogous way to the dimension reduction techniques applied to explicit trait rating data in previous work on the dimensions underpinning first impressions of faces.

Method

Stimuli

The 10k US Adult Faces Database (Bainbridge et al., 2013) contains naturalistic images of 2222 adult face images, captured in unconstrained natural poses. Faces vary in ethnicity (83.7% White, 9.9% Black, 3.2% Hispanic and 3.1% Asian), sex (42.9% female), age, emotional expression and pose.

Procedure and participants

To collect free-text descriptions (i.e., written first impressions) of the images while reducing the potential for participant fatigue, 22 sets of faces (each set consisting of 101 images) were first created by random sampling from the face-image database. Two hundred and forty-four participants (140 women and 100 men, mean age = 40.10 years, SD = 13.14 years) were then randomly allocated to one of these image sets. In each image set, participants were sequentially presented with a face and a text box and were asked to describe their first impressions of that person by entering text into the text box. Participants were also instructed that they could use whatever words came to mind and however many words they wanted. The trial order (i.e., the order in which faces were presented) was fully randomized and the study was self-paced. The study was run online and participants were recruited via Prolific (constrained to the United Kingdom and United States, with English as a first language), and data were collected using the Gorilla online testing platform (Anwyl-Irvine et al., 2020). Each face received free-text descriptions from an average of 10.50 participants (SD = 0.89).

Text processing pipeline

Topic modelling is an unsupervised learning approach that is capable of parsing sets of documents, or a corpus, to automatically cluster together words or phrases that have high semantic coherence (Steyvers

& Griffiths, 2014). Moving from raw text to topic extraction requires pre-processing steps, highlighted here.

First, a single document file for each face was produced by concatenating all descriptions given for that face into a single document or paragraph. Text such as digits and punctuation (for example, exclamation or question marks) were removed, clear misspellings of words were corrected and all text was converted to lowercase. Stop-words were then removed (e.g., function words like 'always', 'because' and 'but'). Each document was then lemmatized – that is, words are replaced with their *lemma*, or the most basic representation of a set of word forms – for example, the words 'feels', 'felt' and 'feeling' have the lemma '*feel*'. Lemmatized words are words removed from cues that help form sentence structure, or offer information about temporal context (e.g. *felt* indicating past tense). Practically, lemmatization reduces variation in text data and makes it easier to capture the meaning of words present in a text. Instead of counting instances of words such as *'smiling'* and *'smiled'* as separate words, the count of the lemma *'smile'* is obtained.

The processed documents were converted from text to a matrix representation using a two-step process. The entire set of lemmatized documents (or corpus) was initially converted into a bag-of-words (BOW) representation. All individual words across every document (here, a description for a face) are isolated, indicating the number of unique words in the corpus. The number of times each word appears within each document (face description) is counted. This creates a sparse matrix representation with documents as rows, columns as words and entries as non-negative counts, capturing the occurrence of words in documents. While this representation discards word order, it has the distinct benefits of simple computation and being directly interpretable – each column is tied to the instance of a word. The BOW representation was then rescaled using the term-frequency inverse-document-frequency (TF-IDF) method. TF-IDF scales the count representation of each word according to how often it appears in each document relative to its appearance in other documents. Higher values in the matrix represent a word that is used more frequently in that document, but not others.

To extract topics, the TF-IDF representation of the corpus was subjected to Non-Negative Matrix Factorization (NMF; Lee & Seung, 1999). NMF takes an observed n x p matrix (n documents, p words), a predefined number of components, and estimates two matrices (H and W), which, when multiplied, return the observed matrix (with some error). The H and W matrices have useful properties for topic modelling, as they naturally cluster together words into topics and assign document affinities for a given topic. For example, for the extraction of three components, NMF would yield H, a p x 3 matrix where higher entries would indicate greater association between words and the component, and W, an n x 3 matrix where higher entries indicate greater association between a document and the components. The components can be given meaning (i.e. topics) by examining the words with the largest entries in H and the representative documents with the highest entries in W.

NMF requires an a priori number of components to extract. Given that this number of components/topics ties directly to the conclusions about the psychological structure of first impressions, we sought to evaluate the quality of the NMF matrix estimations using an objective topic coherence measure from computational linguistics, the UMass score (Mimno et al., 2011). The UMass score computes the log probability of observing a given pair of words in a random document in the corpus, divided by the number of documents, computed across all possible pairings of words in a topic. Scores closer to zero are more coherent (Mimno et al., 2011; Rüdiger et al., 2022). Based on the number of components found in previous ratings-based studies, we evaluated topic coherence in NMF solutions with two to six components. We limited the coherence score to consider the top 20 words of the resulting components, a reasonable default that topic modelling algorithms show stable correlations with human-derived topics from text corpora (Arun et al., 2010; Řehůřek & Sojka, 2010; Röder et al., 2015). The average UMass coherence across components was used to select the final number of components/topics.

The code and data underpinning these analyses are available at the OSF (osf.io/chxnp). Text processing and NMF were carried out with Python and the scikit-learn package (Pedregosa et al., 2011).

Results

Corpus characteristics

Participants, on average, tended to provide relatively short raw text descriptions (mean number of words per description = 2.55 words, *SD* = 2.44, median = 2). After combining responses for a description of each face, each face had descriptions of on average 26.79 words in length (*SD* = 8.68, median = 25). Importantly, NMF has previously been shown to work well for topic modelling with even short text samples (Albalawi et al., 2020). Moreover, a TF-IDF representation of text has also shown good results for topic modelling on short document corpuses, such as social media posts (Lossio-Ventura et al., 2019).

There were a total of 36,006 words extracted across these descriptions, which comprised 4612 unique word entries. Figure 1a illustrates the frequencies of the top 20 words, which shows a concentration of word usage around positive words such as *happy, friendly* and *kind*. This matrix was then subjected to NMF for topic extraction.

Topic modelling

NMF was repeated for two to six topics, and at each iteration, the UMass coherence score was calculated on the 20 words with the highest loading on each topic to establish the most coherent set of word-to-topic loadings. These scores indicated that a model with two topics produced the most coherent fit (two topics UMass = −2.48, three topics UMass = −2.82, four topics UMass = −2.84, five topics UMass = −2.90, six topics UMass = −3.10; see Figure 1a). Split-half cross-validation of this showed strong support for this two-topic solution (see Supplementary Materials), showing cross-validated Ochiai coefficients for Topics 1 and 2 of .85 and .66, respectively.

Figure 1b shows the loadings for the 20 words, with the highest loadings for each topic. The words *'happy', 'friendly', 'kind', 'smile'* and *'nice'* showed the highest loadings on Topic 1, suggesting Topic 1 reflected first impressions related to aspects of positive emotional valence, warmth and approachability. By contrast, the words *'serious', 'sad', 'angry', 'unhappy'* and *'shy'* showed the highest loadings on Topic 2, suggesting Topic 2 reflected first impressions related to negative emotional valence and potential threat (Topic 2). Using the loading each face received on each topic, we carried out a Bayesian correlation (with a flat prior on the correlation coefficient), to estimate the degree of association between each topic, $r = -.65$, 94% Credible Interval [−0.67, −0.62], indicating that a higher loading on Topic 1 suggests a generally lower loading on Topic 2. We created average faces based on the faces with the highest ($n = 100$) and lowest ($n = 100$) loadings on each topic. These visualizations are shown in Figure 1b and illustrate how cues of positive facial affect (e.g. smiling) are associated with Topic 1 and how more neutral and masculine cues are associated with Topic 2.

We also replicated this analysis using a different topic modelling method, Latent Semantic Analysis (LSA), which uses singular-value decomposition to extract topics. This alternative approach has a different structure that forces orthogonal components as well as allowing for positive and negative loadings. This approach also confirmed a two-topic structure, and the resulting topics showed minimal differences to those extracted by NMF (see Supplementary Materials).

Topic modelling conditional on face sex

The two-topic structure described above was generated from both female and male faces. An important possibility to consider is whether a fundamentally different topic structure might emerge when considering female and male faces in isolation. To test this, we repeated the entire analysis above, building a separate corpus for male ($n = 1267$) and female ($n = 955$) faces separately and subjecting those to topic

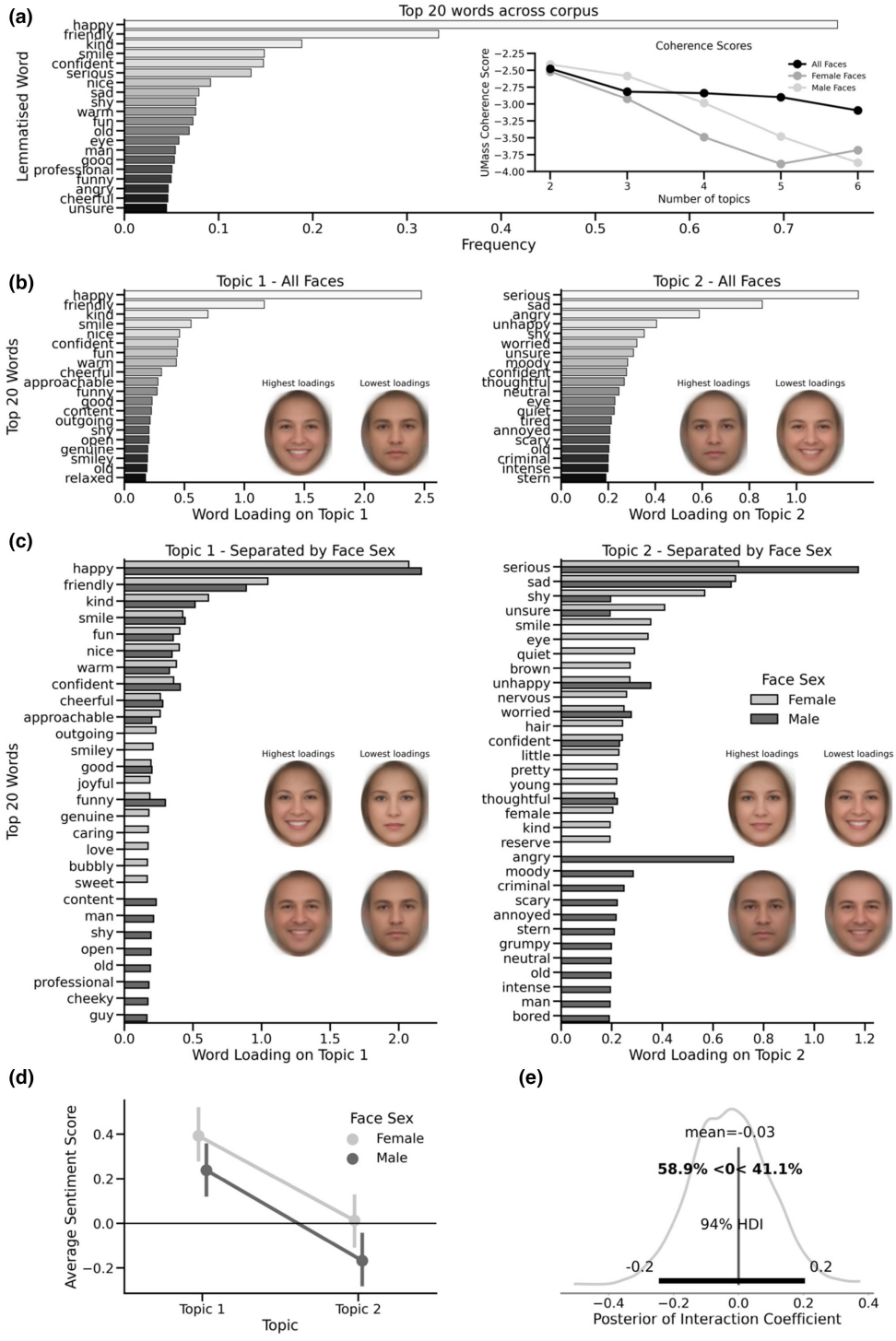

FIGURE 1 Panel (a) shows the 20 most frequent words across the entire corpus. The inset shows coherence scores for candidate topic numbers for the entire dataset and topics within each sex. Panel (b) shows word loadings for each of the two topics extracted from the entire dataset, including composite faces. Panel (c) shows word loadings for two topics extracted from female and male faces separately and the associated facial appearances. Panel (d) shows average sentiment scores and posterior credible intervals within each topic, as derived from each sex. Panel (e) shows the posterior distribution of the interaction coefficient and the probability that males have lower (more negative) sentiment under Topic 2 as compared to females.

modelling. For female faces, a total of 15,309 words were used, with 2640 unique entries. For males, 20,697 words were used in total, with 3457 unique entries.

For both the female and male faces corpora, a two-topic solution emerged consistently (females; two topics UMass = −2.52, three topics UMass = −2.92, males; two topics UMass = −2.41, three topics UMass = −2.59; see Figure 1a). As with the full dataset, the topics derived from analysing descriptions of female and male faces separately showed a negative correlation (female faces: $r = -0.64$ [−0.67, −0.60]; male faces: $r = -0.65$ [−0.68, −0.62]). Topics for male and female faces showed broad similarities, albeit with some clear differences. Figure 1c illustrates the top 20 words loading on each topic for both female and male topic models. For Topic 1, there is broad similarity between the highest-loaded words – for example, the words *happy, friendly, kind, smile* and *fun* had the highest loading for female faces, and *happy, friendly, kind, smile* and *confident* were the highest for males, while only female faces had loadings for words like *love, sweet, caring* and *bubbly*. Conversely, male faces only had positive loadings for words like *professional, content, open* and *shy*. Thus, while Topic 1 seems to broadly represent the same kind of positively-valanced impressions for both sexes, there are some differences between sexes in the language used.

Topic 2 broadly indexed negatively valanced, threat-related impressions. For female faces, the words *serious, sad, eye, shy* and *unsure* had the highest loadings, while for male faces, the words *serious, angry, sad, unhappy* and *moody* had the highest loadings. Notably, there were more unique words associated with each topic for each sex. Female faces had loadings for words like *anxious, reserve, nervous* and *little*, as well as relatively descriptive words such as *hair* and *brown*. Male faces, however, had loadings for words such as *angry, moody, criminal, annoyed, scary, unfriendly, grumpy* and *unapproachable*. While the topics within sexes show the same kind of positive and negative valence impressions, male faces receive a greater variety of negative descriptions. Visualizations of the facial characteristics associated with the topics for male and female faces are shown in Figure 1c.

Sentiment analysis of male and female topic structures

The topic structures for female and male faces are broadly similar but show some qualitative differences in the loading of words used. For example, for Topic 2, a wider variety of negatively valanced words appear to be used to describe male faces. To investigate these differences further, we employed a sentiment analysis of the top 20 highest loading words for each topic, separately for female and male faces. Sentiment analysis uses dictionary-based methods (Hutto & Gilbert, 2015) to score the valence of written text from −1 (extremely negative) to 1 (extremely positive). We submitted the top 20 words of each topic, for each sex, to sentiment analysis and then tested for differences in average sentiment between conditions using a Bayesian linear regression model. Specifically, the model tested the interaction between sex and topic. We coded the sex and topic variables to have the reference values 'female' and 'Topic 1', respectively, and as such, the interaction coefficient indicates the *difference-in-difference* (i.e., whether the average difference in sentiment scores between males and females is greater for males under Topic 2).

The model indicated a clear effect of topic, such that the words loading highly on Topic 2 had a lower sentiment score on average than those in Topic 1, with the 94% credible area of the posterior distribution of the coefficient comfortably excluding zero, $b = -0.38$ [−0.53, −0.22]. That is, Topic 2 sentiment was approximately −0.38 units lower than Topic 1, on average. In addition, the model indicated that the words for the male topic model very likely received lower sentiment on average than the female topic model, $b = -0.15$ [−0.31, 0.01], with a probability of 97% for a negative effect. Male faces received on average −0.15 sentiment units less than female faces, regardless of topic. However, the interaction term showed no clear evidence of males having credibly different sentiment scores for Topic 2 as compared to females, $b = -0.03$ [−0.24, 0.20], with a probability of just 41% of being positive. These results are shown in Figure 1d,e and indicate that the difference in means between Topic 1 and Topic 2 is credibly similar in magnitude for both female and male faces.

Interim discussion

Using natural language processing techniques, we uncovered the emergence of a two-topic structure from written first impressions of a large sample of face images. These topics were generated from spontaneously written first impressions and seem to broadly capture aspects of positive emotional valence, warmth and approachability (Topic 1) and negative emotional valence and potential threat (Topic 2). The two topics were negatively correlated. Splitting the dataset by face sex revealed that a similar two-topic structure emerges for both female and male faces, with some substantial overlap in the key words used for each topic.

STUDY 2 – COMPARISONS OF TEXT AND RATING-BASED MODELS OF FACIAL FIRST IMPRESSIONS

Analysis of free-text descriptions of faces in Study 1 suggested that written descriptions of first impressions are underpinned by two topics that appear to primarily reflect impressions of positive emotional valence, warmth and approachability (Topic 1) and impressions of negative emotional valence and potential threat (Topic 2). As these dimensions are conceptually similar to the Valence and Dominance components previously found to underpin explicit ratings of faces (Jones & Kramer, 2021; McAleer et al., 2014; Morrison et al., 2017; Oosterhof & Todorov, 2008), Study 2 investigated the relationships between the topics derived from free-text descriptions in Study 1 and components derived from PCA of explicit trait ratings of the same face images. Specifically, we instantiated the Valence–Dominance model for these faces by conducting a PCA on ratings of the same traits used in previous studies that generated the valence–dominance model of first impressions (Jones et al., 2021; Oosterhof & Todorov, 2008).

Method

Ratings of the 2222 faces used in Study 1 for the traits *attractive, unhappy, sociable, emotionally stable, mean, boring, aggressive, weird, intelligent, confident, caring, egotistic, responsible* and *trustworthy* (14 of the 15 traits used to derive the valence–dominance model; Oosterhof & Todorov, 2008) were publicly available from their original study (Bainbridge et al., 2013). Because the traits used to derive the valence–dominance model in previous work also included the additional trait of dominance, we recruited 225 additional participants (mean age = 38.94 years, $SD = 14.27$ years, 132 females) who each rated one of 22 subsets of images (101 images per image subset) for dominance using a 1 (not very) to 7 (very) scale (average number of raters per face = 10.23, $SD = 0.42$), with trial order fully randomized. These additional participants (i.e., those who rated the images for dominance) were recruited via Prolific with the same constraints on recruitment as Study 1. Intraclass correlations for these traits varied from 0.26 to 0.44 (see Figure S1).

Following previous studies of the Valence–Dominance model (Jones et al., 2021; Oosterhof & Todorov, 2008), we carried out a PCA on the averaged ratings for each trait per face, retaining the first two components only. We then regressed these components onto the loadings each face received for the two topics obtained in Study One, using Bayesian linear regression and adjusting for the sex of each face.

Results

Valence dominance model extraction

The full correlation structure of the PCA is shown in Table 1. The first principal component explained 59% of the variance in trait ratings and showed high correlations with traits such as *unhappy* ($r = 0.91$),

TABLE 1 Loadings of individual traits with each component of the valence–dominance model.

	Valence–dominance model	
Trait	**PC1**	**PC2**
Unhappy	**0.91**	0.02
Caring	**−0.91**	−0.22
Sociable	**−0.89**	0
Attractive	**−0.6**	0.36
Confident	**−0.69**	0.57
Dominance	0.02	**0.69**
Mean	**0.90**	0.25
Intelligent	**−0.74**	0.27
Weird	**0.76**	−0.24
Trustworthy	**−0.9**	−0.12
Egotistic	**0.61**	**0.61**
Boring	**0.56**	−0.34
Responsible	**−0.84**	0.05
Aggressive	**0.84**	0.36
Emotionally Stable	**−0.86**	0.02

Note: Values above |.5| are highlighted in bold.

caring ($r = -0.91$), *sociable* ($r = -0.89$) and *trustworthy* ($r = -0.90$), while the second component explained 12% of the variance and was highly correlated with *dominance* ($r = 0.69$) and *egotistic* ($r = 0.61$). This pattern of results aligns closely with those reported in previous studies of the valence–dominance model (64% and 18% variance explained in the first and second components, respectively, Oosterhof & Todorov, 2008). As such, we refer to the first component as Valence and the second as Dominance. Examination of the rest of the principal components showed that the third principal component explained 8.1% of the variance (like the 6% explained by the third component found in the initial valence–dominance model extraction; Oosterhof & Todorov, 2008). Beyond this, each component explained less than 4.2% of the successive variance. Taken together, the first and second components captured 71% of the variance in the trait ratings data.

Our PCA was calculated using a singular value decomposition of the correlation matrix between the 15 trait ratings and then ordering the singular values in decreasing order. No rotations were carried out (i.e., principal components retain their maximal-variance properties).

Mapping the valence dominance model to the topic model

Since the direction of principal components is arbitrary, we rescaled the first PC (Valence) so that higher values represented more positive valence (i.e., altering only the sign of the correlations stated above). We z-score standardized all variables beforehand, except for face sex (coded one for male), and fit a model that predicted topic loadings with no intercept, the main effect of each principal component (PC1 = Valence, PC2 = Dominance), the main effect of each topic (whether a loading belonged to Topic 1 or Topic 2), and the interactions between each principal component and each topic. Face sex was included as a covariate.

By standardizing all variables and removing the intercept, we can directly interpret the coefficients as standard deviation unit changes in topic loadings with a 1SD increase in loadings on the principal component. We used weakly regularizing normal priors on each coefficient (mean zero and standard deviation one, Gelman et al., 2020) and a t-distributed likelihood to be robust to the influence of outliers (Kruschke, 2014).

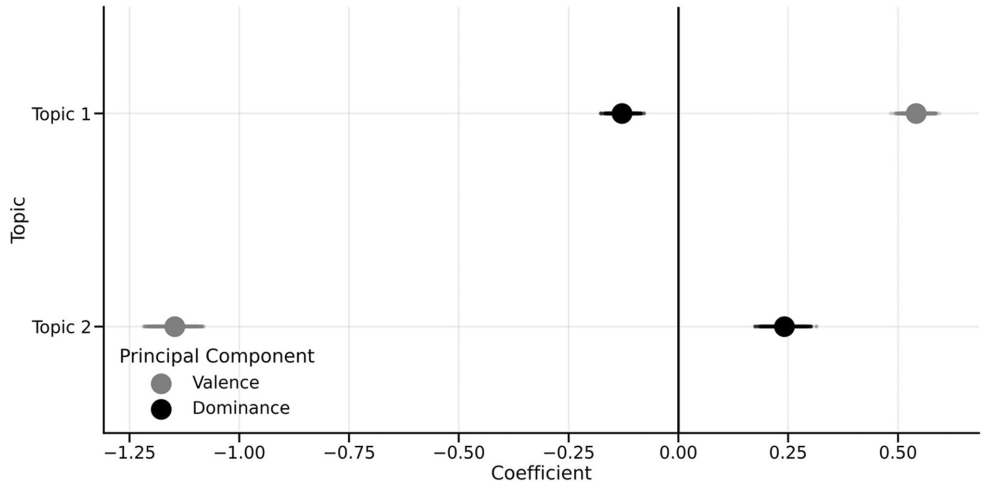

FIGURE 2 Coefficients and 94% credible intervals for the association between each principal component and the loadings on each topic.

Adjusted for sex, the valence–dominance model explained 44.4% [43.2, 45.7] of the variance in topic loadings. Examining the coefficients showed clear and strong associations between each component and each topic loading. For Topic 1, increases in PC1 (Valence) were associated with positive loadings, $b = 0.54$ [0.52, 0.57], and increases in PC2 (Dominance) were associated with lower loadings, $b = -0.13$ [−0.16, −0.10]. For Topic 2, increases in PC1 were associated with much lower loadings, $b = -1.15$ [−1.19, −1.11], while PC2 showed a smaller but positive relationship, $b = 0.24$ [0.21, 0.28]. The posterior distributions of each coefficient comfortably excluded zero and ranged in size from approximately one tenth of an SD change in topic loadings (PC2's association with Topic 1) through to just over one SD unit (PC1's association with Topic 2). These results indicate clear associations between the valence–dominance model and the topic loading model. However, under half of the variance in the topic loadings was accounted for by the valence–dominance model. Coefficients are shown in Figure 2.

GENERAL DISCUSSION

In Study 1, topic modelling of free-text descriptions of participants' first impressions of 2222 face images revealed that these impressions were underpinned by two topics (i.e., dimensions), reflecting perceptions of positive emotional valence, warmth and approachability (Topic 1) and perceptions of negative emotional valence and potential threat (Topic 2). This pattern of results suggests that free-text descriptions of first impressions contain consistent statistical regularities that can be extracted using natural language processing (NLP) methods. Importantly, the same pattern of results was observed when free-text descriptions of all images were analysed and when descriptions of male and female images were analysed separately, suggesting that the pattern of results is not simply a by-product of differences in how participants described male and female faces.

Studies that derive models of first impressions by analysing explicit trait ratings of face images have typically reported that first impressions are underpinned by Valence and Dominance components that are either orthogonal or weakly correlated (e.g., Jones et al., 2021; Oosterhof & Todorov, 2008). In Study 2, we replicated this pattern of results in our analyses of explicit trait ratings. However, our analyses in Study 2 also suggested that there is a potentially important quantitative distinction between models derived from these two types of data. Collectively, the Valence and Dominance components derived from analyses of explicit trait ratings of face images explained ~44% of the variance in the topics extracted

from free-text descriptions. This suggests that, although there is some overlap in the dimensions underpinning first impressions derived from these two types of responses, models derived from explicit ratings do not fully capture the language that participants typically use to describe first impressions. Existing models of first impressions are non-redundant with topic models, and natural language-based approaches highlight many aspects of first impressions that are unrelated to personality, as indicated by the range of words related to emotional states and physical appearance. As such, natural language offers an alternative approach to studying first impressions and offers support to the current literature, as well as novel insights into the complexity of the process. Moreover, it offers a 'high fidelity' method of studying first impressions and similar approaches, being able to capture rich information about processes as they are spontaneously produced, as opposed to forcing observers to make judgements of specific traits.

That the Valence and Dominance components derived from analyses of explicit trait ratings of face images explained ~44% of the variance in the topics extracted from free-text descriptions may have relatively straightforward explanations. First, the traits used to estimate the Valence–Dominance model based on Oosterhof and Todorov's (2008) approach were based on standardized images. As naturally varying images were used here, these traits may not capture the impressions attributed to these more variable images. However, we closely reproduce the Valence–Dominance model in this data, and it is unlikely the model would appear only under standardized conditions. An alternative explanation is that the traits used to generate the Valence–Dominance are already filtered down to just personality-related first impressions, and not first impressions in general. For example, some words captured by the topic model include physical descriptors (*old, smiley*) and emotional states (*angry, sad*), which go beyond the content of the Valence–Dominance model, and as such, that model may be unable to capture information about general impressions present in text data. While the data-driven approach could in principle be expanded to include non-personality-related words by collecting ratings, it would require an increase in the number of traits to be rated by additional participants. Working directly with text data captures aspects of first impression that are ignored in traditional approaches.

However, the difference can also be explained by a fundamental distinction between 'can' and 'do' approaches in psychology (Satchell et al., 2022). Explicit trait ratings are clearly useful approximations of first impressions, but a significant limitation is that they 'coerce' participants into considering a judgement of a trait that they may not have spontaneously generated. To elaborate, observers can make ratings of traits like *emotional stability* when explicitly prompted to rate stimuli on that trait, but such trait ratings may not map straightforwardly to the topics generated from free text, which constitute the impressions that observers form in the moment they view the face. Thus, text-based approaches may offer a new route to advance models of social perception (Mondloch et al., 2023), as they can capture actually-formed impressions moments after that impression is formed. By contrast with previous research that has used text-based methods to generate a list of traits for further observers to rate (Oosterhof & Todorov, 2008; Sutherland et al., 2018), more directly analysing text responses themselves also avoids the assumption that a different set of participants would necessarily spontaneously generate similar responses. By using natural language, a complex process such as first impressions can be modelled more fully, and in doing so we find that various features that are unrelated to personality judgements are relatively common – social impressions encompass more than just personality judgements.

The extracted topic models show broad theoretical similarities to existing models of social perception. For example, the words people use to describe their impressions of others can be distilled into a smaller number of features (i.e., topics), a feature of many social cognitive models outside of face perception. The stereotype content model – underpinned by the dimensions of warmth and competence or the 'Big Two' axes of social cognition (Fiske, 2018; Fiske et al., 2007) – emerges across varied domains of psychological science and reflects core aspects of functioning in a social world (Martin & Slepian, 2021). More specifically, these two dimensions often appear as variations around concepts such as agency and communality (Eagly, 1997), particularly behaviours focused on self-interest, promotion and preservation, or community and prosociality, respectively (Ybarra et al., 2008). The content of Topic 1 aligns closely with the latter dimension, with language describing smiling, positive and prosocial appearances. The content of Topic 2 aligns particularly with agentic concepts and the identification of

individuals who pose threats or harm or are displaying negative emotions. The topics also lend support to the recent idea that the fundamental axes of social cognition reflect a functional, gendered perception of the social world (Martin & Slepian, 2021). Though a two-topic solution emerged for both female and male faces when considered separately, there were divergences and overlaps between the words used for each of those topics. For example, in Topic 2, the words *anxious* and *nervous* were exclusively used for female faces, while male faces were described using words such as *angry* and *scary*, which suggests the kinds of positive and negative valence perceived are adjusted to some extent conditionally on the sex of the observed individual.

The results also point to an update in our understanding of models of facial first impressions. Traditional data-driven models based on ratings posit aspects of valence and dominance as separate components that may be orthogonal (Oosterhof & Todorov, 2008; Sutherland et al., 2013) or cast other impressions such as competence and warmth into separate components (Lin et al., 2021). With the high-fidelity nature of text data, the derived topic model here suggests that positive and negative valence emerge as separate dimensions on their own but are strongly correlated. This pattern differs from existing models in that it suggests valence is a primary 'thread' that ties first impressions together in a low-dimensional space. These results align with a general and fundamental 'approach or avoid' continuum that underpins perception, governing whether a stimulus or conspecific should be moved towards or away from (Allport, 1935; Bamford & Ward, 2008; Chen & Bargh, 1999; Jones & Kramer, 2021). Similar findings have emerged using clustering methods that, based on trait ratings, partition faces into a fundamental 'approach or avoid' decision (Jones & Kramer, 2021). The topic model here may suggest that what underpins first impressions is an evaluation that leads to a binary decision about whether to approach or avoid a conspecific.

In the current work, we deliberately did not constrain free-text responses in any way. A limitation of this approach was that, without instruction, participants tended to produce relatively short free-text descriptions of individuals. Thus, while our analyses suggest that models derived from unconstrained free-text descriptions and explicit trait ratings show conceptual – but nonredundant – overlap, it remains an open question whether more constrained free-text descriptions would produce topics identical to those seen here. Systematically varying the amount of text participants are required to write in each description would be needed to clarify this issue. Varying the context in which people are asked to generate free-text descriptions (e.g., instructing participants that the individuals shown are job applicants versus potential romantic partners) would also allow the generalizability of the topics across assessment contexts to be probed.

To conclude, we show for the first time that the statistical regularities in free-text responses of facial first impressions are underpinned by two topics that reflect perceptions of positive emotional valence, warmth and approachability (Topic 1) and negative emotional valence and potential threat (Topic 2). Components derived from explicit trait ratings of face images explain ~44% of the variance in the topics extracted from free-text descriptions, demonstrating that models built from free-text responses are non-redundant with existing models of first impressions derived that rely on ratings and demonstrating the utility of natural language as a vehicle for more closely studying first impressions. The derived topic model here suggests a subtly different psychological organization of facial first impressions, suggesting separate components for both positive and negative valence, which subsume separate or orthogonal components from existing approaches. Future work can compare how natural language-based models and traditional data-driven approaches can explain impressions in varying social contexts or whether the blending of these models can offer a greater understanding of first impressions. More generally, a key avenue of investigation is how natural language processing can advance and refine models of social cognition.

AUTHOR CONTRIBUTIONS

Alex L. Jones: Conceptualization; investigation; writing – original draft; methodology; visualization; writing – review and editing; software; formal analysis; data curation. **Victor Shiramizu:** Investigation; writing – review and editing; resources; data curation. **Benedict C. Jones:**

Investigation; funding acquisition; writing – original draft; writing – review and editing; validation; project administration.

CONFLICT OF INTEREST STATEMENT

All authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at https://osf.io/chxnp/.

ORCID

Alex L. Jones https://orcid.org/0000-0003-3600-3644

REFERENCES

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *42*(3), 1–4. https://doi.org/10.3389/frai.2020.00042

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (pp. 798–844). Clark University Press.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in knowledge discovery and data mining* (pp. 391–402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Bamford, S., & Ward, R. (2008). Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion*, *8*(2), 174–183. https://doi.org/10.1037/1528-3542.8.2.174

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706. https://doi.org/10.1016/j.jrp.2009.03.007

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Ding, K., Choo, W. C., Ng, K. Y., & Zhang, Q. (2023). Exploring changes in guest preferences for Airbnb accommodation with different levels of sharing and prices: Using structural topic model. *Frontiers in Psychology*, *14*, 1–5. https://doi.org/10.3389/fpsyg.2023.1120845

Dumas, R., & Testé, B. (2006). The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology*, *65*(4), 237–244. https://doi.org/10.1024/1421-0185.65.4.237

Eagly, A. H. (1997). Sex differences in social behavior: Comparing social role theory and evolutionary psychology. *American Psychologist*, *52*(12), 1380–1383. https://doi.org/10.1037/0003-066X.52.12.1380.b

Eggleston, A., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Culturally learned first impressions occur rapidly and automatically and emerge early in development. *Developmental Science*, *24*(2), e13021. https://doi.org/10.1111/desc.13021

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fruhen, L. S., Watkins, C. D., & Jones, B. C. (2015). Perceptions of facial dominance, trustworthiness and attractiveness predict managerial pay awards in experimental tasks. *The Leadership Quarterly*, *26*(6), 1005–1016. https://doi.org/10.1016/j.leaqua.2015.07.001

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories (1st edition)*. Cambridge University Press.

Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2017). Facial appearance affects science communication. *Proceedings of the National Academy of Sciences*, *114*(23), 5970–5975. https://doi.org/10.1073/pnas.1620542114

Hester, N., Jones, B. C., & Hehman, E. (2021). Perceived femininity and masculinity contribute independently to facial impressions. *Journal of Experimental Psychology: General*, *150*(6), 1147–1164. https://doi.org/10.1037/xge0000989

Holzleitner, I. J., Lee, A. J., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., DeBruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1589–1595. https://doi.org/10.1037/xhp0000685

Hutto, C. J., & Gilbert, E. (2015). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Jack, R. E., Crivelli, C., & Wheatley, T. (2018). Data-driven methods to diversify knowledge of human psychology. *Trends in Cognitive Sciences*, *22*(1), 1–5. https://doi.org/10.1016/j.tics.2017.10.002

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2022). Understanding the role of faces in person perception: Increased reliance on facial appearance when judging sociability. *Journal of Experimental Social Psychology*, *100*, 104288. https://doi.org/10.1016/j.jesp.2022.104288

Jaeger, B., Sleegers, W. W. A., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, *75*, 102125. https://doi.org/10.1016/j.joep.2018.11.004

Jaeger, B., Wagemans, F. M. A., Evans, A. M., & van Beest, I. (2018). Effects of facial skin smoothness and blemishes on trait impressions. *Perception*, *47*(6), 608–625. https://doi.org/10.1177/0301006618767258

Jones, A. L., & Jaeger, B. (2019). Biological bases of beauty revisited: The effect of symmetry, averageness, and sexual dimorphism on female facial attractiveness. *Symmetry*, *11*(2), 279. https://doi.org/10.3390/sym11020279

Jones, A. L., & Kramer, R. S. S. (2021). Facial first impressions form two clusters representing approach-avoidance. *Cognitive Psychology*, *126*, 101387. https://doi.org/10.1016/j.cogpsych.2021.101387

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Kleisner, K., Chvátalová, V., & Flegr, J. (2014). Perceived intelligence is associated with measured intelligence in men but not women. *PLoS One*, *9*(3), e81237. https://doi.org/10.1371/journal.pone.0081237

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423. https://doi.org/10.1037/0033-2909.126.3.390

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lefevre, C. E., & Lewis, G. J. (2014). Perceiving aggression from facial structure: Further evidence for A positive association with facial width–to–height ratio and masculinity, but not for moderation by self–reported dominance. *European Journal of Personality*, *28*(6), 530–537. https://doi.org/10.1002/per.1942

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications*, *12*(1), 5168. https://doi.org/10.1038/s41467-021-25500-y

Liu, S., Zhang, R.-Y., & Kishimoto, T. (2021). Analysis and prospect of clinical psychology based on topic models: Hot research topics and scientific trends in the latest decades. *Psychology, Health & Medicine*, *26*(4), 395–407. https://doi.org/10.1080/13548506.2020.1738019

Lossio-Ventura, J. A., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2019). Clustering and topic modeling over tweets: A comparison over a health dataset. *Proceedings IEEE International Conference on Bioinformatics and Biomedicine*, *2019*, 1544–1547. https://doi.org/10.1109/bibm47256.2019.8983167

Martin, A. E., & Slepian, M. L. (2021). The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition. *Perspectives on Psychological Science*, *16*(6), 1143–1158. https://doi.org/10.1177/1745691620904961

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

Mehu, M., Little, A. C., & Dunbar, R. I. M. (2007). Duchenne smiles and the perception of generosity and sociability in faces. *Journal of Evolutionary Psychology*, *5*(1), 183–196. https://doi.org/10.1556/jep.2007.1011

Menegatti, M., Pireddu, S., Crocetti, E., Moscatelli, S., & Rubini, M. (2021). The Ginevra de' Benci effect: Competence, morality, and attractiveness inferred from faces predict hiring decisions for women. *Frontiers in Psychology*, *12*, 658424. https://doi.org/10.3389/fpsyg.2021.658424

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. Conference on Empirical Methods in Natural Language Processing. https://www.semanticscholar.org/paper/Optimizing-Semantic-Coherence-in-Topic-Models-Mimno-Wallach/ef2d64e448ee5ed2dc26179c8570803ded123a5e

Mondloch, C. J., Twele, A. C., & Thierry, S. M. (2023). We need to move beyond rating scales, white faces and adult perceivers: Invited commentary on Sutherland & Young (2022), understanding trait impressions from faces. *British Journal of Psychology*, *114*(2), 504–507. https://doi.org/10.1111/bjop.12619

Morrison, D., Wang, H., Hahn, A. C., Jones, B. C., & DeBruine, L. M. (2017). Predicting the reward value of faces and bodies from social perception. *PLoS One*, *12*(9), e0185093. https://doi.org/10.1371/journal.pone.0185093

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions from faces. *Psychological Science*, *30*(1), 65–79. https://doi.org/10.1177/0956797618813092

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110. https://doi.org/10.1007/s10919-009-0082-1

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly*, *25*(5), 817–834. https://doi.org/10.1016/j.leaqua.2014.06.002

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Olivola, C. Y., Sussman, A. B., Tsetsos, K., Kang, O. E., & Todorov, A. (2012). Republicans prefer republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science*, *3*(5), 605–613. https://doi.org/10.1177/1948550611432770

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Quist, M. C., Watkins, C. D., Smith, F. G., DeBruine, L. M., & Jones, B. C. (2011). Facial masculinity is a cue to women's dominance. *Personality and Individual Differences*, *50*(7), 1089–1093. https://doi.org/10.1016/j.paid.2011.01.032

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*(1), 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399–408. https://doi.org/10.1145/2684822.2685324

Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLoS One*, *17*(4), e0266325. https://doi.org/10.1371/journal.pone.0266325

Satchell, L., Jaeger, B., Jones, A., Lopez, B., & Schild, C. (2022). *Beyond reliability in first impressions research: Considering validity and the need to "mix it up with folks"*. PsyArXiv. https://doi.org/10.31234/osf.io/4gk6b

Shiramizu, V. K. M., Lee, A. J., Altenburg, D., Feinberg, D. R., & Jones, B. C. (2022). The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports*, *12*(1), 22479. https://doi.org/10.1038/s41598-022-27124-8

South Palomares, J. K., & Young, A. W. (2018). Facial first impressions of partner preference traits: Trustworthiness, status, and attractiveness. *Social Psychological and Personality Science*, *9*(8), 990–1000. https://doi.org/10.1177/1948550617732388

Steyvers, M., & Griffiths, T. (2014). Probabilistic topic models. In *Handbook of latent semantic analysis*. Routledge. https://doi.org/10.4324/9780203936399.ch21

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology*, *113*(4), 1056–1078. https://doi.org/10.1111/bjop.12583

Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, *44*(4), 521–537. https://doi.org/10.1177/0146167217744194

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

Talamas, S. N., Mavor, K. I., Axelsson, J., Sundelin, T., & Perrett, D. I. (2016). Eyelid-openness and mouth curvature influence perceived intelligence beyond attractiveness. *Journal of Experimental Psychology: General*, *145*(5), 603–620. https://doi.org/10.1037/xge0000152

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences*, *1124*(1), 208–224. https://doi.org/10.1196/annals.1440.012

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, *25*(3), 395–410. https://doi.org/10.1080/02643290802044996

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. https://doi.org/10.1037/a0032335

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Tzschaschel, E., Brooks, K. R., & Stephen, I. D. (2022). The valence-dominance model applies to body perception. *Royal Society Open Science*, *9*(9), 220594. https://doi.org/10.1098/rsos.220594

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Ybarra, O., Chan, E., Park, H., Burnstein, E., Monin, B., & Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology*, *38*(7), 1083–1092. https://doi.org/10.1002/ejsp.559

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497–1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jones, A. L., Shiramizu, V., & Jones, B. C. (2024). Decoding the language of first impressions: Comparing models of first impressions of faces derived from free-text descriptions and trait ratings. *British Journal of Psychology*, *00*, 1–16. https://doi.org/10.1111/bjop.12717