Call for Research on the Impact of Information Retrieval on Social Norms

Tim Gollub[1], Pierre Achkar[2], Martin Potthast[3], and Benno Stein[1]

[1] Bauhaus-Universität Weimar
[2] Leipzig University, Fraunhofer ISI Leipzig
[3] Kassel University, hessian.AI, ScaDS.AI

Abstract. The information retrieval (IR) systems of major media platforms have a significant impact on social norms. Social norms contribute to the cultural identity of a society, but can also lead to people being marginalized, suffering from social pressure, and feeling inferior. For this reason, we call on the IR community to (1) contribute to the social sciences with computational means to study the impact of IR systems on social norms, and (2) to incorporate respective social science research findings into IR system development. To support our call, this paper presents a dataset and classification technology for investigating the prevalence of normative beauty ideals in multimodal (image and text) search results and recommendations. On a dataset comprising 928 annotated social media posts, in addition to determining the best classification model for the task, we examine how state-of-the-art zero-shot classifiers perform compared to fine-tuned models, and how multimodal models perform compared to unimodal variants. With 92% classification accuracy, a late fusion model with individually fine-tuned image and text representations achieves peak effectiveness, suggesting that this technology is mature enough to be used in computational social science research and IR systems. To illustrate our work, we analyze the image search results pages of a major Web search engine and report our findings. The code repository of our research is available at https://github.com/webis-de/ecir25-beauty-ideals.

Keywords: computational social science · social norms · norm-beauty · multimodal classification · ethical IR

1 Introduction

The ranking and recommendation algorithms of information retrieval (IR) systems play a central role in determining the media content users encounter online. From search results pages to the personalized content displayed on social media feeds, IR systems meanwhile shape a significant portion of our daily digital experiences. Social science research shows that media consumption has a profound impact on social norms and ideals [1, 4, 16, 19]. By selectively presenting specific types of content, IR systems influence individuals' perceptions of reality, including concepts related to success, happiness, relationships, and personal

Fig. 1. Illustrative samples of "norm-beauty" (left) and "non-norm-beauty" (right) images generated using Stable Diffusion XL.

identity. Given this influence, it is apparent that IR systems not only mediate access to information but also impact social norms. Social norms, once shaped, affect individual and collective well-being. For example, normative beauty ideals disseminated through online media have been linked to increased rates of depression, body dissatisfaction, and self-esteem issues [13, 16, 18]. In this context, it becomes evident that IR systems, which shape the content seen by billions, also indirectly shape mental health outcomes and the broader well-being of society.

Despite the great social influence of IR systems, their ranking and recommendation algorithms are largely optimized from a commercial perspective. Content that generates a high level of engagement and therefore higher advertising revenue is given preferential treatment. The social impact of these algorithms is rarely considered (a notable exception being the "safe search"' filters). However, given the wide reach of these systems, it is vital that greater consideration is given to social impact in the design and development of IR systems. To enable more effective research into the social impact of IR algorithms (see Section 2 for the current state of the art), we call on the IR community to participate in the development of tools for large-scale computational analysis of media content with respect to socially meaningful constructs such as beauty, status, or values. Recent advances in natural language processing and computer vision offer promising opportunities in this regard.

To this end, in this paper, we focus on the concept of norm-beauty. Our research question is to what extent modern machine learning models can learn whether a social media post showing a human body reflects a Western ideal of beauty. We collaborate with a team of anthropologists and compile an annotated dataset of 928 social media posts. Each post contains an image with a human body (image modality), the user's caption (text modality) and a classification of

the post into one of the classes "norm-beauty"' and "non-norm-beauty". To give an impression of the image modality while preserving privacy, Figure 1 shows eight generated images that are similar to images from the dataset. The dataset is presented in detail in Section 3. We use our dataset to train and evaluate a set of five different norm beauty classification models, which we present in Section 4. Of all the models considered, the best effectiveness is achieved with a multimodal late fusion model, which achieves a classification accuracy of 92% on a hold-out test set. The results of all models are presented in detail in Section 5. As a demonstration of our research, we apply our classifier to Google image search results. By varying the search query with respect to the person's age, ethnicity, and gender, we examine the ratio of norm-beauty image search results across these dimensions. We find that the ratio varies widely, from 6.12% for "adult Indigenous females" to 83.67% for "young Asian females".

However, the most important finding of our research is that current machine learning technology is advanced enough to work with subjective and complex social constructs such as norm-beauty, advocating our call upon the IR community to support the social sciences in the pursuit of socially responsible algorithms.

2 Related Work

This section provides a brief overview of recent social science literature examining the impact of online media on normative beauty ideals, as well as the work in the field of computational aesthetics.

2.1 Online Media, Beauty Ideals, Well-Being Issues

The (negative) effects of online media on normative ideals of beauty are primarily investigated in gender studies, as young women seem to be the most affected [11, 20, 25]. As a theoretical framework, self-discrepancy theory [17], proposing that individuals hold self-perceptions about the actual, ideal, and ought self is often considered. If discrepancies arise between these perceptions, i.e., if the attributes we believe we have don't match the attributes we aspire or we think we should have, this can lead to negative emotions and cognitions. There are numerous studies that show that online media increase the discrepancy between the actual and the ideal self, and that the availability of smartphone apps to alter images exacerbates this effect (cf.[16, 23]).

So far, there are few approaches to possible countermeasures. For instance, efforts to signal edited or digitally altered images, such as labeling images to indicate modification, have not effectively reduced the psychological impact on users. Paradoxically, research suggests that such labels may even heighten awareness of these ideals, reinforcing rather than diminishing their influence [2, 14]. More promising seem approaches of diversification, where positive effects have been reported with so called body-positivity posts on Instagram [8, 20].

A general issue of existing work are small sample sizes, and the question of representativeness is often discussed as a limitation. The studies are either

qualitative interviews (26 young women [25], 200 mainly young women [11]), or quantitative on manually labeled datasets (246 Instagram posts [20], 640 Instagram posts [8]). This situation prompts us to call on the IR community to provide their expertise in developing classification technologies to enable research on more representative data samples.

2.2 Computational Aesthetics

In recent years, interest in computational aesthetics has grown, with researchers increasingly using image classification models to assess or categorize the aesthetic value of images. A comprehensive overview of the field is provided by Bodini, who examines various existing methods and their development from a philosophical and neuroaesthetic perspective [3]. A critical analysis of various datasets is performed, such as the AVA dataset, and their impact on computational aesthetics is evaluated. The paper also discusses the challenges and limitations of aesthetic evaluation, such as the binary criteria of "ugly vs. beautiful". The work points out the complexity of beauty classification and its evaluation, as well as the need to consider broader factors such as cultural and socio-demographic contexts. We take this aspect into account by defining norm-beauty as the subjective perception of specific groups with a common cultural and socio-demographic background. Different reference groups may have different opinions on norm-beauty.

In the digital photography context, Suchecki and Trzcinski approached aesthetic evaluation using a CNN, analyzing a dataset of 1.7 million Flickr photos [26]. By fine-tuning the AlexNet neural network for binary classification, the researchers were able to classify images as aesthetically pleasing or not pleasing with an accuracy of 70.9% in their study. This method, which is based purely on visual information, shows the potential of machine learning to determine aesthetic values in photos, but also the difficulty of the task. Their work provide interesting insights into the features that contribute to the aesthetic appeal of a photograph, such as color saturation, sharpness, and contrast.

The paper "Deep learning for assessing the aesthetics of professional photographs" by Chambe et al. evaluates the effectiveness of aesthetic assessment models in the context of professional photography [6]. The models were initially trained on photographs from the AVA dataset, which is known for its wide variety of aesthetic scores and semantic labels. However, the models encountered new challenges when applied to other types of photography, such as fashion, architecture, and sports. After fine-tuning the models on data with various photographic categories, the results improved significantly. The results of the study underline the positive effect of fine-tuning models to domain-specific data to improve accuracy and reliability. In our work, we are also interested in evaluating the effectiveness gain of fine-tuning models compared to zero-shot models.

The study by Choudhary and Gandhi deals with another area of aesthetics, namely the evaluation of the beauty of faces [7]. The study uses a number of machine learning models to classify the degree of facial attractiveness using the SCUT-FBP dataset, which contains images of Asian women's faces. The classification was carried out both binary (attractive or not attractive) and multi-level

Table 1. Original Instagram captions and the reformulations generated by GPT-4.

Caption	Reformulated caption
Pregnancy era 🫛 @riotswim	Embracing the pregnancy journey with Riot Swimwear.
Woman. 💫	Celebrating womanhood.
Spring's leg 💜☂️🦄♓⏫	Embracing the magical and rejuvenating energy of spring, symbolized by shades of purple and feminine mystique.
Is it bikini season yet? 🥹	Eagerly anticipating the arrival of bikini season.
selfie mood 🦋 #mykonos	Feeling whimsical and free-spirited in my selfie from Mykonos.
life goal: live most days in a swimsuit 🌞	Aspiring to spend the majority of my days basking in the sunshine, clad in my favorite swimsuit.
what a night 🖤	Had an unforgettable night.

(five levels of attractiveness). The best model achieves an accuracy of 88% for the binary classification. More recently, Bougourzi et al. applies an advanced deep learning method for facial beauty prediction (FBP) using the SCUT-FBP5500 dataset, which consists of 5,500 frontal face images with different attributes such as age, gender, and ethnicity [5]. Each of these images was rated by 60 different annotators on a beauty scale of one to five. The REX-INCEP framework is the core element of their approach and combines the capabilities of the ResneXt-50 and Inception-v3 models to optimize feature extraction for FBP. As presented in the next section, our understanding of norm-beauty goes beyond facial/body attractiveness, and takes into account the whole composition of a post, including the image scenery and the text modality of the post.

3 Dataset

This section presents the norm-beauty dataset used for our evaluation. The dataset was compiled as part of our research collaboration with the Institute of Historical and Cultural Anthropology at Tübingen University, Germany.[4]

The dataset consists of 928 social media posts from Instagram, each featuring an image with a person and a caption (the first user comment). To comply with common research practice, only public posts have been taken into account. The annotation of the dataset was done by five young female German university students, who participated in a digital anthropology seminar. Before annotating the posts into the classes "norm-beauty" and "non-norm-beauty", a shared understanding of beauty ideals and an annotation guideline had been developed.

[4] https://www.digitalesbild.gwi.uni-muenchen.de/en/curating-the-feed-interdisciplinary-perspectives-on-digital-image-feeds-and-their-curatorial-assemblages/

It is clear that beauty ideals are neither objective, static, nor universal but rather culturally constructed and mediated notions of (physical) beauty and affective encounters that are shaped by internalized conventions, habitual dispositions, and tacit knowledge learned and acquired over the course of a lifetime. This is why, instead of attempting to represent objective criteria of beauty, our guidelines emulate the annotators' subjective perception of beauty ideals. To annotate, attention has to be paid to the entire composition of the images, the staging and (re-)presentation of the person(s) depicted, the background, the pose, indications for post-processing and the message conveyed in the caption. After observing tacit cultural rules and aesthetic regimes on Instagram, and given the positionality of the annotators (Western culture), the following working definition of (Western) norm-beauty has been formulated: The body and face of the persons depicted are at the center of the image, the representations of bodies tend to reproduce the gender binary by accentuating gender-specific attributes, the bodies are staged as desirable, aspirational and in part sexualized. The representations correspond to Eurocentric ideals of beauty, depicting predominantly White, young and tall, slender and toned, hairless and blemish-free bodies. Furthermore, the posts are characterized by images with beautiful landscapes in the background, soft lightning mood, or by captions that emphasize the attractiveness of the person in the image, or hashtags that highlight bodily beauty (e.g. 'sexy').

The annotation process resulted in a balanced dataset comprising 472 "norm-beauty" and 456 "non-norm-beauty" posts. As part of our data pre-processing, we normalize image colors to RGB and expand the dataset using several image augmentation techniques such as rotation, horizontal flipping, and brightness adjustments. In addition, the original captions are reformulated using GPT-4. Figure 1 shows examples of the original captions and their revised versions.

Furthermore, we transform the image modality into rich text descriptions using the multimodal LLaVA model, providing a detailed narrative for each image. An example of such a transformation for the top right image in Figure 1 reads as follows:

The person in the image is standing with a straight posture, looking directly at the camera. The facial expression is neutral, and there is no overt emotion or attitude conveyed. The body is displayed prominently, with the top half of the suit jacket open, revealing the chest area. The setting is a plain background, which puts the focus on the person. The skin texture appears to be smooth, and there are no visible tattoos or adornments. The person has a full beard and mustache, which are well-groomed. The person appears to be of average weight, with a slim build. There are no visible muscular features, and the facial characteristics are typical of a middle-aged adult. There are no indications of any disabilities or syndromes in the image.

For the experiments presented below, the dataset is divided into training, validation, and test set with a ratio of 80%, 10%, and 10% respectively.

4 Models

This section details the models used for classifying norm-beauty in social media posts, categorized by their modality (vision, language, multimodal) and training approach (zero-shot vs. fine-tuned).

4.1 Vision Models

The image modality is clearly the main modality for our task, and we are interested in how well the classification of norm-beauty can be performed using only the image. From the vast amount of pre-trained vision models, we aim to evaluate (1) popular models, for which strong effectiveness scores have been reported across different tasks, (2) cutting-edge models, which might benefit from recent technological innovations, and (3) domain-specific models, which might benefit from a domain-specific pre-training. Overall, we experiment with the following four image processing models.

Residual Neural Network (ResNet): Utilizes skip connections within a deep convolutional framework to maintain gradient flow and enable effective training of deep layers, enhancing image recognition capabilities [15].

Vision Transformer (ViT): Applies a transformer architecture to images by dividing them into patches, which are processed as sequences, facilitating a global understanding of visual context. Its multi-head self-attention mechanism allows it to capture complex relationships between image segments [10].

Shifted Window Transformer (Swin): Introduces a hierarchical vision structure and shifted windowing mechanism to improve computational efficiency and scalability when processing high-resolution images. This method enables localized attention within windows while maintaining cross-window interactions for feature refinement [22].

Self-Supervised Model (SEER): Trained through self-supervised learning on a vast dataset of Instagram images. Its architectural design enables it to extract detailed and complex visual features and generalize effectively across a range of scenarios, demonstrating robust adaptability and fairness. [12].

All models are fine-tuned on our training set for classifying 'norm-beauty' and 'non-norm-beauty'.

4.2 Language Models

For text classification, we utilize a BERT model due to its robust contextual encoding capabilities [9]. The model is fine-tuned with different transformations: (1) Raw captions, (2) Reformulated English captions with clarified hashtags and emojis, and (3) Image descriptions generated with LLaVA to integrate visual context.

4.3 Multimodal Model

Given the unimodal models for norm-beauty classification of social media posts, we are interested in exploiting both modalities for the classification. A straightforward and widely used approach to combine the two modalities is via so called late fusion. Initially, ViT and BERT are fine-tuned separately on image and text data, respectively. Their outputs are then concatenated into a unified feature vector. This combined vector undergoes further processing in a fully connected layer to produce the final classification.

4.4 Zero-Shot Models

Since, as the experiments presented in Section 5 reveal, the multimodal late fusion model achieves convincing effectiveness, we are interested in the extent to which zero-shot approaches are able to perform norm-beauty classification as well. We experiment with two models, a vision-only CLIP model and the multimodal LLaVA model.

Contrastive Language-Image Pre-training (CLIP): CLIP represents texts and images in a shared latent space, trained via extensive self-supervised learning. The model highlights the efficacy of dual-encoder frameworks in image classification [24]. We use CLIP as a simple baseline zero-shot classifier by measuring the cosine similarity between the CLIP representation of the image modality and CLIP representations for the two classification labels.

Large Language and Vision Assistant (LLaVA): More sophisticated than CLIP, LLaVA merges cutting-edge language and vision technologies to process multimodal content. Developed on a dataset formulated for intricate multimodal chats, LLaVA's architecture, combining CLIP's vision encoder with a Llama-based language model, allows the model to handle complex visual and textual inputs effectively [21].

4.5 Zero-Shot Prompt Engineering

To effectively utilize the LLaVA model for classifying images into 'norm-beauty' or 'non-norm-beauty', we develop a strategic prompt engineering process. Initial tests with basic prompts revealed limitations, leading to inconsistent results. To address this, we designed prompts that direct the model's attention to key visual and contextual features relevant to norm-beauty.

These refined prompts instruct the model to analyze elements such as pose, skin texture, body presentation, and alignment with conventional beauty standards. This structured approach significantly enhances LLaVA's ability to evaluate images comprehensively.

For multimodal analysis, prompts are adjusted to incorporate image captions, enabling the model to integrate visual and textual data effectively. Additionally, a specialized system prompt is designed to set the analytical context for LLaVA, framing it as a social scientist tasked with evaluating social media images through the lens of societal norms. This ensures that the model maintained a consistent focus on diversity and representation in its analyses.

Table 2. Effectiveness comparison of the image classification models on the raw and augmented datasets (DS).

Model	DS	Acc	Prec	Rec	F1
SEER_384	raw	**0.87**	0.87	0.87	**0.87**
SEER_384	aug	0.86	0.85	0.87	0.86
ResNet_384	raw	0.59	0.60	0.57	0.58
ResNet_384	aug	**0.78**	0.78	0.81	**0.79**
ViT_384	raw	**0.84**	0.82	0.87	**0.84**
ViT_384	aug	0.84	0.82	0.87	0.84
Swin_384	raw	**0.84**	0.81	0.89	**0.85**
Swin_384	aug	0.84	0.83	0.85	0.84

Table 3. Effectiveness comparison of the text classification model using the original captions, reformulated captions, and generated descriptions as training dataset (DS).

DS	Acc	Prec	Rec	F1
original	0.74	0.76	0.71	0.74
reformulated	0.76	**0.80**	0.71	0.75
generated	**0.79**	0.73	**0.94**	**0.82**

5 Evaluation

This section reports on the effectiveness of the classification models presented in Section 4, and presents the results of a first showcase study.

5.1 Image Classification

For the uni-modal classification with the four vision models, we evaluate different image resolutions (not reported) and the impact of using augmented images on the effectiveness. The results for the vision models is summarized in Table 2. The domain-specific SEER models, particularly SEER_384 on raw data, achieve the best effectiveness scores. However, they tend to overfit when trained on augmented data. ResNet models exhibit a varied behavior in response to changes in resolution and data augmentation. ViT and Swin Transformer models yield high recall values, especially on the raw data, but face potential overfitting on the augmented data.

5.2 Text Classification

Three types of text were considered as input for the BERT classifier: original captions, reformulated captions, and detailed image descriptions. This approach allowed for a thorough analysis of BERT's classification abilities across various textual forms. Fine-tuning BERT was carefully calibrated, with a learning rate of 1×10^{-5} across two epochs.

The effectiveness of the model across different textual forms is presented in Table 3. It is interesting to observe that the original captions provided a reasonable benchmark, with considerable accuracy and F1 score, underlining BERT's ability to process informal and heterogeneous text. Still, a significant improvement was observed with reformulated captions, confirming the importance of pre-processing for improved clarity and coherence. However, the highest level of effectiveness was achieved through detailed descriptions, emphasizing the significance of context-rich textual data in enhancing classification accuracy.

Table 4. Effectiveness comparison of the Late Fusion model with Base and Fine-Tuned representations for the two modalities.

Model	Acc	Prec	Rec	F1
vit384_bert_base	0.88	0.89	0.87	0.88
vit384_bert_fine_tuned	**0.92**	**0.91**	**0.93**	**0.92**

5.3 Multimdodal Classification

By developing a multimodal classification framework that combines visual and textual data, the aim is to uncover the complexity of social media posts to understand whether there is a complementary relationship between the text and the images in the dataset.

For the Late Fusion model, the results summarized in Table 4 demonstrate the effects of using base versus fine-tuned versions of the ViT and BERT models within the Late Fusion architecture. The aim is to show how these two configurations influence the effectiveness of the model in classifying the data. The Late Fusion model, which combines the strengths of ViT and BERT, shows a significant improvement in classification effectiveness compared to using each model individually. This improvement therefore emphasizes the complementary nature of the visual and textual information in the data to enable a more complete understanding of the task.

Altogether, both configurations have acceptable values for training and validation losses, indicating their ability to learn and generalize adequately. Remarkably, the fine-tuned configuration shows a strong increase in test effectiveness, with test accuracy and F1 score significantly better than the base configuration. This demonstrates the fine-tuned model's ability to better learn from and classify the data, and points to the benefits of using models that are optimized for the given data.

5.4 Zero-Shot Classification

Two zero-shot classification models have been tested. The first one is CLIP, which has shown promising capabilities in zero-shot classification by using large image-text pairs to enable robust learning of visual representations. As noted by Zhang et al. [27], this innovative approach can transfer knowledge effectively across different tasks without additional training. The second one is LLaVA 1.5, which is known for both its remarkable effectiveness on various benchmarks and as versatile zero-shot classifier capable of integrating multi-modal data through advanced natural language prompts. The effectiveness of both models on the test dataset is summarized in Table 5.

The effectiveness of the CLIP model is close to random, which underlines the complexity of the task. CLIP's method, designed to correlate images with text, may not fully capture the subtleties required for this challenge. When evaluating the effectiveness of the LLaVA model in different experimental configurations,

Table 5. Effectiveness comparison of the Zero-Shot models using the vision only and vision+text modalities.

Model	Type	Acc	Prec	Rec	F1
Clip_336	Vision	0.49	0.50	0.83	0.62
llava-1	Vision	0.67	0.61	**1.00**	0.76
llava-2	Vision	0.75	0.75	0.77	0.76
llava-1	Vision + Text	0.73	0.65	**1.00**	0.79
llava-2	Vision + Text	**0.85**	**0.84**	0.87	**0.85**

a systematic analysis shows the different effects of prompt detail, integration of multi-modal data (images and text), and the application of user-defined system prompts. This structured investigation aims to clarify the factors that contribute to the model's ability to correctly classify the data.

The initial set of experiments, which focused on vision-only analysis, illustrates the importance of prompt complexity. The use of a more general prompt (llava-1) resulted in the model achieving a moderate level of accuracy (0.67) and F1 (0.76). In contrast, the use of a more detailed prompt (llava-2) significantly improved the model's effectiveness, which showed an increase in accuracy (0.75). This improvement supports the assumption that a comprehensive prompt enables a more accurate interpretation by the model and thus improves its classification capability.

Next, the analysis was extended to include textual data in addition to visual input. When captions were added to prompt llava-1, the model showed an improvement in both accuracy (0.73) and F1 (0.79). The addition of captions to llava-2 also results in a considerable increase in effectiveness, achieving an accuracy of (0.85) and the highest F1 score of (0.85). This demonstrates the positive effect of adding textual context as an extra input to improve the model's ability to accurately recognize and classify content.

5.5 Showcase

To illustrate the application domain of our classifiers, we conduct an experiment to determine the fraction of images retrieved from Google search that fall into the "norm-beauty" category. We use the work of Lazuka et al. to guide the attributes used to formulate the queries [20]. These attributes are age group (young, adult, middle-aged), gender (male, female), and ethnicity (African American/Black, Asian, Caucasian, Indigenous, Latinx, Middle Eastern), resulting in 36 unique query combinations. Each query is executed in the Google Chrome browser in incognito mode from a computer in Europe, and the first 50 images from each search are retrieved. Using the pre-trained ViT_384 classifier, we analyze the images and calculate the percentage of images classified as "norm-beauty". The results are given in Figure 2. With 83.67%, the highest fraction of norm-beauty images is observed for the query "young Asian female", followed by 76.00% for

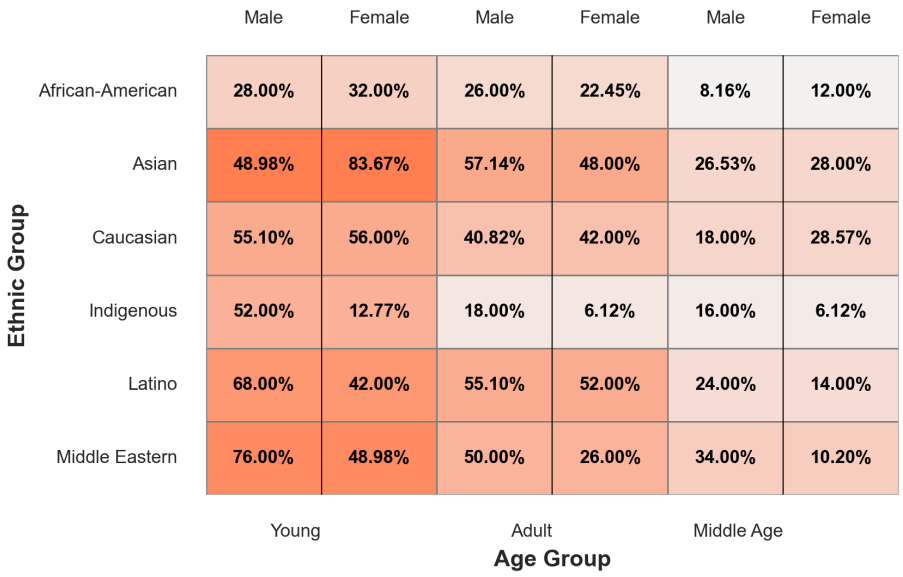

Fig. 2. Results of the showcase study: Percentage of norm-beauty images in the image search results obtained for queries that differ by stated Ethnicity, Gender, and Age.

the query "young Middle Eastern male". In general, it is striking how different the fractions are across the queries. Since the experiment is meant as a showcase, we refrain from interpreting the results, but rather point out the apparent potential of using such models in the context of social science research.

6 Conclusion

As IR systems increasingly shape our online experiences, the importance of designing these systems with social impacts in mind cannot be overstated. This paper demonstrates the feasibility of using current technological advancements to analyze complex sociological constructs within media content. Our Late Fusion model strategically combines the strength of both image and text modalities, and its strong effectiveness underlines our call for collaborative computational social science research. By bridging the gap between IR technology and social science research, we can begin to develop algorithms that not only serve commercial interests but also contribute positively to societal well-being and healthier social norms.

Acknowledgements This research is funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) – project number 421299207.

Bibliography

[1] Arias, E.: How does media influence social norms? experimental evidence on the role of common knowledge. Political Science Research and Methods **7**(3), 561–578 (2019)

[2] Bissell, K.L.: Skinny like you: Visual literacy, digital manipulation and young women's drive to be thin. Studies in media & information literacy education **6**(1), 1–14 (2006)

[3] Bodini, M.: Will the machine like your image? automatic assessment of beauty in images with machine learning techniques. Inventions **4**(3) (2019), https://doi.org/10.3390/inventions4030034

[4] Borg, K.: Media and social norms: Exploring the relationship between media and plastic avoidance social norms. Environmental Communication **16**(3), 371–387 (2022)

[5] Bougourzi, F., Dornaika, F., Taleb-Ahmed, A.: Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. Knowl. Based Syst. **242**, 108246 (2022), https://doi.org/10.1016/J.KNOSYS.2022.108246

[6] Chambe, M., Cozot, R., Meur, O.L.: Deep learning for assessing the aesthetics of professional photographs. Comput. Animat. Virtual Worlds **33**(6) (2022), https://doi.org/10.1002/CAV.2105

[7] Choudhary, G., Gandhi, T.K.: Indexing facial attractiveness and well beings using machine learning. In: 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 1–6 (2016), https://doi.org/10.1109/R10-HTC.2016.7906813

[8] Cohen, R., Irwin, L., Newton-John, T., Slater, A.: # bodypositivity: A content analysis of body positive accounts on instagram. Body image **29**, 47–57 (2019)

[9] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), URL http://arxiv.org/abs/1810.04805

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net (2021), URL https://openreview.net/forum?id=YicbFdNTTy

[11] Gill, R.: Being watched and feeling judged on social media. Feminist Media Studies **21**(8), 1387–1392 (2021), https://doi.org/10.1080/14680777.2021.1996427

[12] Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision. CoRR **abs/2202.08360** (2022), URL https://arxiv.org/abs/2202.08360

[13] Grabe, S., Ward, L.M., Hyde, J.S.: The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. Psychological bulletin **134**(3), 460 (2008)
[14] Harrison, K., Hefner, V.: Virtually perfect: Image retouching and adolescent body image. Media Psychology **17**(2), 134–153 (2014)
[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), URL http://arxiv.org/abs/1512.03385
[16] Henriques, M., Patnaik, D.: Social Media and Its Effects on Beauty. IntechOpen (May 2021), ISBN 9781839624483, https://doi.org/10.5772/intechopen.93322
[17] Higgins, E.T.: Self-discrepancy: a theory relating self and affect. Psychological review **94**(3), 319 (1987)
[18] Keles, B., McCrae, N., Grealish, A.: A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. International journal of adolescence and youth **25**(1), 79–93 (2020)
[19] Kryston, K., Eden, A.: I like what you like: Social norms and media enjoyment. Mass Communication and Society **25**(5), 603–625 (2022)
[20] Lazuka, R.F., Wick, M.R., Keel, P.K., Harriger, J.A.: Are we there yet? progress in depicting diverse images of beauty in instagram's body positivity movement. Body Image **34**, 85–93 (2020), ISSN 1740-1445, https://doi.org/https://doi.org/10.1016/j.bodyim.2020.05.001
[21] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. CoRR **abs/2304.08485** (2023), https://doi.org/10.48550/ARXIV.2304.08485
[22] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CoRR **abs/2103.14030** (2021), URL https://arxiv.org/abs/2103.14030
[23] MacCallum, F., Widdows, H.: Altered images: Understanding the influence of unrealistic images and beauty aspirations. Health Care Analysis **26**(3), 235–245 (Jul 2016), ISSN 1573-3394, https://doi.org/10.1007/s10728-016-0327-1
[24] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. CoRR **abs/2103.00020** (2021), URL https://arxiv.org/abs/2103.00020
[25] Siddiqui, A.: Social media and its role in amplifying a certain idea of beauty. Infotheca **21**(1), 73–85 (2021), ISSN 2217-9461, https://doi.org/10.18485/infotheca.2021.21.1.4
[26] Suchecki, M., Trzcinski, T.: Understanding aesthetics in photography using deep convolutional neural networks. CoRR **abs/1707.08985** (2017), URL http://arxiv.org/abs/1707.08985
[27] Zhang, R., Wei, Z., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of CLIP for few-shot classification. CoRR **abs/2207.09519** (2022), https://doi.org/10.48550/ARXIV.2207.09519