Original Manuscript

Social Science Computer Review
2024, Vol. 0(0) 1–24
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393241269394
journals.sagepub.com/home/ssc
Sage

Forty Thousand Fake Twitter Profiles: A Computational Framework for the Visual Analysis of Social Media Propaganda

Noel George[1], Azhar Sham[1], Thanvi Ajith[1], and Marco Bastos[1,2]

Abstract

Successful disinformation campaigns depend on the availability of fake social media profiles used for coordinated inauthentic behavior with networks of false accounts including bots, trolls, and sock-puppets. This study presents a scalable and unsupervised framework to identify visual elements in user profiles strategically exploited in nearly 60 influence operations, including camera angle, photo composition, gender, and race, but also more context-dependent categories like sensuality and emotion. We leverage Google's Teachable Machine and the DeepFace Library to classify fake user accounts in the Twitter Moderation Research Consortium database, a large repository of social media accounts linked to foreign influence operations. We discuss the performance of these classifiers against manually coded data and their applicability in large-scale data analysis. The proposed framework demonstrates promising results for the identification of fake online profiles used in influence operations and by the cottage industry specialized in crafting desirable online personas.

Keywords

propaganda, teachable machine, DeepFace, romance scam, Twitter moderation research consortium

Introduction

Influence operations orchestrated by the Kremlin-linked Internet Research Agency (IRA) to target the 2016 US presidential election created the blueprint for large-scale disinformation campaigns leveraging social media platforms to seed division and explore socially contentious issues (Bastos & Farkas, 2019; Freelon et al., 2020). This playbook is based on the creation and management of

[1]University College Dublin, School of Information and Communication Studies, Belfield, Ireland
[2]Department of Media, Culture and Creative Industries, City, University of London, London, UK

Corresponding Author:
Marco Bastos, School of Information and Communication Studies, University College Dublin, Newman Building, Room C121, Belfield, Dublin 4, Ireland.
Email: marco.bastos@ucd.ie

fake social media profiles to push divisive narratives, including conspiracy theories that are impervious to mitigation strategies like fact-checking (Stencel & Luther, 2021). The supply chain of fake social media profiles is also central to luring individuals into fraudulent schemes, including the cottage industry of pig-butchering scams specialized in crafting desirable online personas.

Social media platforms have devised community guidelines to prevent the spread of false or misleading information about elections and civic processes. Twitter's Civic Integrity policy lists misleading information about how to participate in an election or other civic process, false or misleading information that causes confusion or undermines public confidence in an election or other civic process, unverified information about election rigging, ballot tampering, vote tallying, or certification of election results, inciting violence or illegal behavior to interfere with an election or other civic process, and coordinated reporting, posting, or sharing of information to manipulate the public conversation (Twitter, 2021).

Upon identifying and attributing the source and target of such campaigns, social media platforms label, remove, or reduce the visibility of such content depending on the severity and reach of the violation. Very Large Online Platforms (VLOPs) also take action against accounts that repeatedly violate their policies, such as suspending or permanently banning the accounts. The monitoring and identification of such activities relies on users who report content in violation of the Terms of Service. Social media platforms also work with cybersecurity companies and a host of organizations that monitor and verify information related to elections and civic processes.

Within Twitter, these initiatives would eventually mature into the Twitter Moderation Research Consortium (TMRC). Starting in 2018 as a reaction to the large influence operation carried out by the IRA during the 2016 US Presidential Election (Gadde & Roth, 2018; Twitter, 2019), Twitter's then-head of Trust and Safety, Del Harvey, oversaw the company's efforts to safeguard elections and deal with problematic content that could jeopardize healthy conversations. Data would be shared with the academic community in 2017–2018, initially under the umbrella of Twitter's Elections Integrity initiative, which identified and ultimately removed false accounts, Twitterbots, and sockpuppets operated by the Internet Research Agency. The first data release included 2752 accounts the company attributed to the IRA. This list was expanded in early 2018 to include 3814 IRA-linked accounts (Harvey & Roth, 2018). In the following years, the TMRC would grow to include 115,474 unique Twitter accounts removed from the platform due to breaches of the Terms of Service. These accounts posted in excess of 100 million tweets linked to 57 separate influence operations.

The TMRC was the most comprehensive database of social media propaganda available to researchers. It stood out in providing the profile photos and the totality of the content posted by the fake accounts. This is critical for research on social media propaganda, which leverages visual markers to entrust veracity and trigger emotional response, as visual content is often assumed to be a candid reflection of reality (Powell et al., 2015). Research on visual disinformation, however, remains relatively forthcoming, especially in comparison to the large body of work dedicated to textual disinformation (Brennen et al., 2021; Garimella & Eckles, 2020). In the following we address this gap by describing and testing a set of open-source models trained to identify visual markers of coordinated inauthentic behavior conspicuous in the TMRC database and ultimately adopted by the romance scam industry. The codebase for the models detailed in this study is provided with the training and test datasets, and we expect this material to be relevant for research in inauthentic online activity and mitigation strategies.

Previous Work

Successful propaganda campaigns on social media are supported by the streamlined and cost-effective creation of fake social media profiles that also marked a shift in the frames and compositional choices employed by propagandists. Social media propaganda has leveraged the affordances of social platforms and quickly supplanted the typical militaristic tropes of state propaganda by subtler, more insidious penetration tactics across communication networks (Arif et al., 2018). Digital tools such as Twitterbots, fake accounts, sock puppets, trolls, and compensated influencers have become the instruments of modern propaganda campaigns that circulate at scale owning to the socially embedded, impactful, and creative visual composition of such campaigns (Bastos & Mercea, 2019; Benkler et al., 2018).

Visual propaganda is increasingly employed by state and non-state actors operating vast networks of fake accounts on social media. These actors meticulously manipulate user profile images to influence and engage culturally and politically entrenched audiences (Bastos et al., 2023). Previous research has expounded the profound impact of emotional and human-centric narratives on social media platforms, underscoring the effectiveness of strategies that exploit the vulnerability of human perceptions to biases (Seo, 2019), including racial prejudice promptly weaponized for political objectives (Freelon et al., 2020). Profile pictures play a significant role in visual misinformation, serving as initial touchpoints on social platforms and acting as visual representations of individuals and groups. These images are curated to motivate trust, evoke emotions, or resonate with specific lifestyles. Indeed, the inherent ability of images to command attention and evoke strong emotional responses renders them invaluable for persuasion (Iyer et al., 2014; von Sikorski, 2022). Propagandists have aptly utilized visual media to steer public opinion, exploiting cognitive biases and eliciting emotions to advance their agendas (Qian et al., 2022).

Such visual mechanisms are critical to how information is perceived (Peng et al., 2023). Visual features such as realism or aesthetic appeal can lend undue credibility to misinformation, exploiting the very heuristics that influence audience perception. This is particularly relevant for social media profile pictures, with the presence of human faces on social media profiles being associated with higher user engagement, which in turn is conducive to trustworthiness and likability (Li & Xie, 2020; Sasahara et al., 2020). Consequently, the choice of a profile picture is more than a personal expression; it can serve as a powerful tool for influencing public perception by exploiting the complex interplay between knowledge, trust, and narrative alignment that underpins credibility. Audiences with limited knowledge about a topic are particularly susceptible to the persuasive power of information sources that appear credible, as the trustworthiness and expertise attributed to these sources can compensate for the audience's limited knowledge (Khatib, 1989).

Previous research has charted how the visual strategies employed by the IRA displayed cultural insight and familiarity with the social identity of their targets by trafficking in the tropes of ordinary urban men and attractive young women. This body of work identified differences across campaign targets, with males more likely to appear in the BlackLivesMatter and Russian targeted groups, and females dominating the profile composition of Christians, Conservatives, and particularly Trump supporters (Bastos et al., 2023; Freelon et al., 2020). BlackLivesMatter activists and Christians were more likely to be depicted with high angles, whereas low angles prevailed among Trump supporters and Russian groups. Selfies and self-portraits prevailed in the profiles of BlackLivesMatter activists and Conservatives, while Christians and Trump supporters were framed with regular close-ups. Amateur profile photos were frequent, except for Trump supporters and Christian profiles, which featured a higher incidence of professional profile photos featuring sensually crafted images of young women typical of soft advertisements employed by the cosmetic industry (Bastos et al., 2023).

As such, social media profile pictures extend beyond simple visual identifiers as they embody users' identities and often contain critical information that is not explicitly stated in the profile bio. Liu et al. (2016) found that the profile picture choices of over 66,000 Twitter users varied along with their personality traits, and highlighted their potential for use in misinformation campaigns. A prime example that speaks to the potential of visual communication for social media propaganda was the seminal influence operation during the 2016 US presidential election, when the IRA manipulated profile pictures on Twitter to disseminate divisive social issues, reportedly influencing public opinion and effecting change to the process of democratic deliberation (Benkler et al., 2018). It manufactured credibility by manipulating social media profiles, creating sock-puppet accounts that mimic the behavior of genuine users, buying followers or likes to simulate popularity, and adopting the racial and ideological characteristics that resonate with the target communities (Freelon et al., 2020). Such activities contribute to the illusion of authenticity and can effectively mask the intentions of influence operations.

Fake Twitter profiles often adhere to distinct visual patterns or frames to maximize the odds of a successful group infiltration. These often include the exploitation of the female body to elicit emotional responses from the audience. The IRA in particular has expertly crafted credible and relatable online personas by devising profile pictures that appeal to different demographic groups, with the most salient features in the profile image composition encompassing social dimensions like gender, race, emotion, sensuality, and photographic framing among others (see Appendix 1 for the codebook, including the coded variables). This study is informed by this body of research and attendant literature on visual misinformation, including studies on deepfakes (Dobber et al., 2021; Hameleers et al., 2020; Sundar et al., 2021; Vaccari & Chadwick, 2020), bias in visual communication (Peng et al., 2023), and the narrower literature on audience's perception of photographic images, often mistakenly assumed to represent 'unfiltered reality' (Hameleers et al., 2020; Li & Xie, 2020; Sundar, 2008). While the literature on this topic is largely restricted to the manual coding of images, and is therefore of limited scalability, we seek to leverage Google's Teachable Machine and the DeepFace library to build state-of-the-art classifiers that can meet this challenge.

Methods

While services based on large language models like GPT-4V and Gemini Pro offer powerful capabilities in image description, their generalist nature, lack of transparency, inherent bias, and high resource requirements hinder their application in scientific research where precision, specialization, and ethical considerations are paramount. While such models are suited for Natural Language Processing tasks, open-source models like TensorFlow and Keras are best suited for reproducible research in computer vision due to their flexibility, accessibility, performance, cost-effectiveness, and community support. The customizable capabilities, fine-tuning resources, cost-effectiveness, end-to-end workflows from data collection to preprocessing, and transparency of pre-trained open-source models are particularly important for research leveraging deep learning and neural networks for tasks such as image classification, object detection, and image segmentation.

Indeed, recent advancements in computer vision techniques have facilitated the extraction of visual attributes like face orientation and eye movement, which would otherwise be labor-intensive to classify manually (Peng et al., 2023), and allowed for combining traditional content analysis of visual formats and machine learning algorithms to process large datasets (Lu & Pan, 2022; Matz et al., 2019; Peng, 2018; Peng & Jemmott III, 2018; Talamas et al., 2016). We leverage these advancements to implement a two-phased approach in the classification of the TMRC database. The first phase relies on Google's Teachable Machine to classify the images based on subjective criteria such as sensuality, but also objective technical indicators such as the type of photo and camera angle. The second phase relies on the DeepFace Python library to automate the classification of images by gender and race while also assessing the precision and reliability of the library.

Teachable Machine is a machine learning tool that offers an intuitive interface for model training and deployment. It enables users to easily train models using a variety of drag-and-drop inputs like images, sounds, and poses. The platform operates by collecting and processing labeled examples through deep learning algorithms to identify patterns in the data. Teachable Machine's base model MobileNet was trained on a vast dataset with several categories and the knowledge acquired during its foundational training can be repurposed through transfer learning to recognize new user-defined classes (Figure 1), thereby requiring minimal data and training time for new models (Carney et al., 2020). These models can be customized by adjusting parameters like epochs, learning rate, and batch size to alter the length and speed of the training cycle and to improve precision and recall while preventing overfitting and overshooting. Once trained (Zhang & Peng, 2022), these models can be deployed elsewhere using TensorFlow.

The model training entailed two phases. The first used Teachable Machine for the classification of sensual images. Initially set to 50 epochs, the performance evaluation suggested that accuracy would improve by increasing the number of epochs without affecting the batch size and learning rate. Key performance metrics such as accuracy per class, confusion matrix, accuracy per epoch, and loss per epoch were monitored. An increase to 83 epochs resulted in sensual image classification accuracy improving from .92 to .98, while non-sensual image accuracy declined slightly. Accuracy started to plateau above 90 epochs, suggesting overfitting and the 83-epoch model as the most effective, which was then exported to Keras for downstream analysis using Python. Benchmarking for the sensuality category was performed against the 'racy' category in Google Vision API, which was trained on ImageNet to identify suggestive or provocative elements, including minimal clothing, evocative poses, or close-ups of sensitive body parts without being explicitly adult in nature (Szegedy et al., 2015).

Six other models were trained using MobileNet and PoseNet architectures to classify camera angles and types of shots, with a standard epoch count of 50 and a batch size of 16. Learning rates were set at .001 for MobileNet and .0001 for PoseNet. Manual validation with random images and subsequent error feedback refined the models without altering the initial parameters to maintain training reliability. The outcome of these sessions resulted in the optimal epochs for each model after adjustments for performance: CAD1-MN at 60 epochs, CAD1-PN at 62, CAD2-MN at 50, CAD2-PN at 50, TSD-MN at 80, and TSD-PN at 60. These adjustments achieved a balance between low loss and high accuracy, with CAD1 models requiring additional epochs and CAD2 models achieving near-perfect results.

The other models were trained using the DeepFace library, which is optimized for face recognition and performing detailed facial attribute analysis (Serengil & Ozpinar, 2020).

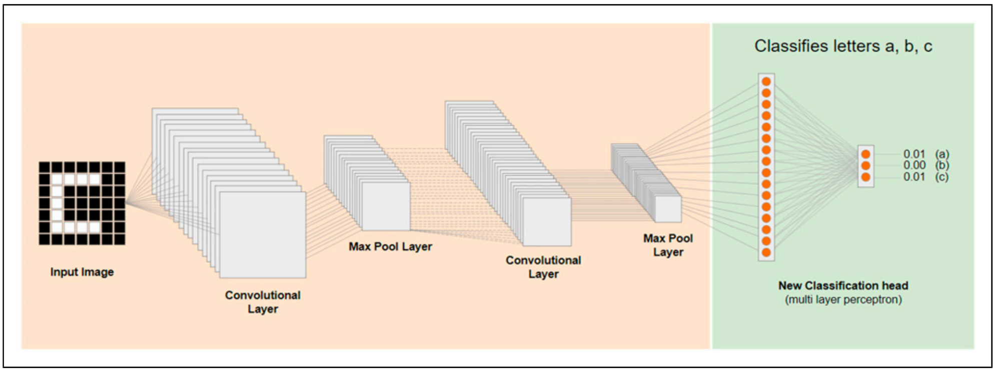

Figure 1. Teachable Machine's transfer learning model (Google, 2017).

DeepFace incorporates various state-of-the-art face recognition models, including VGG-Face, which was trained on 2.6M images of 2.6K individuals, but also FaceNet, OpenFace, DeepID, Dlib, and ArcFace. DeepFace's features cover initial facial feature detection and alignment to representation and verification through multiple facial detectors, including OpenCL, OpenCV, SSD, MTCNN, DLIB, and RetinaFace. Key to our analysis is the Multi-Task Convolution Neural Network (MTCNN), a three-stage cascade framework for face detection and facial landmarks localization that we leverage as the primary facial recognition system owing to its high accuracy (Zhang et al., 2020). Multi-Task Convolution Neural Network's three interconnected deep learning models, Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net) progressively improve the accuracy of face detection by identifying potential face areas in the image, refining the scan area to reduce false positives, and outputting precise face locations and landmarks (Zhang et al., 2020).

DeepFace includes a specialized module for detailed facial attribute analysis integrated with the face recognition module. This module excels at predicting complex attributes like age, gender, emotion, and race. Facial landmarks, such as those around the eyes, nose, mouth, and jawline are crucial for machine learning models to accurately determine facial structure and orientation. They also provide essential information about facial structure and orientation and enable effective face normalization, especially by measuring the distance between the eyes and the positions of mouth corners for emotion analysis (Çeliktutan et al., 2013). Gender and race identification is based on deep learning models pretrained on extensive datasets, particularly the VGG-Face model. These models use facial landmarks identified by MTCNN in conjunction with deep learning pattern analysis to detect and analyze attributes like gender, emotion, and race. The VGG-Face model features an accuracy of 97% for gender and 68% for race prediction.

Framework

We build on the categories mapped in the body of work reviewed above to devise an analytical framework based on DeepFace and Teachable Machine that can identify the compositional tropes of visual propaganda. By processing and analyzing the TMRC database, we identify social and technical indicators in the composition of these images, including camera placement, zoom angles, and distance, which are indicative of the intention to distort one's perceptions of the subject (Huang et al., 2002). Camera angles can be used to subtly shape perceptions and power dynamics, while the type of shot typically influences the viewer's emotional connection to the subject (Wang & Cheong, 2009). This machine learning framework offers the possibility of rapid scalability and support for the analysis of inauthentic behavior exploited in political propaganda and online scams that manipulate one's emotional response (Qian et al., 2022).

This framework relies on computer vision, a specialized branch of computer science focused on enabling machines to interpret digital images, a set of techniques that can bridge the gap between traditional research methods and artificial intelligence (Joo & Steinert-Threlkeld, 2018; Torres & Cantú, 2022; Williams et al., 2020). In addition to the variables mapped in the research discussed above, we also identify sociodemographic variables such as gender and race that also influence the perception of credibility (Khatib, 1989). As discussed in the previous sections, social bots and sockpuppets carefully emulate cultural and sociodemographic traits to maximize perceived in-group association and credibility within the target group. This perceived membership, crafted through mimicry and strategic alignment, fosters intergroup trust that is critical to influence operations masquerading as trustworthy sources (Marwick & Boyd, 2011; Metaxas & Mustafaraj, 2010; Zimmer & Proferes, 2014). In the following, we detail the categories comprising this framework for the study of inauthentic visual communication.

Gender and Race

As a social construct, gender encompasses the expected roles, behaviors, and activities assigned to individuals based on societal norms (Butler, 1999; West & Zimmerman, 1987). Gender norms guide the personalization strategies employed by real and fake social media profiles alike, with the visual representation of gender in profile images often reinforcing societal stereotypes whereby women are objectified and depicted in a sexualized manner (Bastos et al., 2023; Davis, 2018). Gender is thus a central dimension in the creation of fake profiles and the construction of digital identities (Muscanell & Guadagno, 2012; Toma & Hancock, 2012), with 'catfishing' epitomizing the centrality of such social constructs to craft persuasive and convincing personas that mimic real-life stories and engage in communication styles that adhere to stereotypical gender expectations.

Another meaningful social category is race (Lewontin, 1972), which operates as a powerful stratifier and is central to the persistence of inequalities (Bonilla-Silva, 1997). Race also shapes cultural norms and traditions by intersecting with intergroup interaction and intrinsic biases, stereotypes, and prejudice (Omi & Winant, 2014). Race is thus central to identity formation, intergroup relationships, political participation, and cultural acceptance (Bonilla-Silva, 1997). Racial categorizations intersect with other categories to suppress or wield structures that legitimize power imbalances and inequalities (Crenshaw, 1997), including of course the intersection of race and gender, which further compounds their influence on individual experiences (Smedley & Smedley, 2005). The intersection between gender and race, furthermore, is particularly challenging for machine learning algorithms to evaluate (Buolamwini & Gebru, 2018).

Sensuality

Sensuality is a more difficult social vector to define due to its experiential (subjective) and physical (tangible) dimensions. But the distinction between sensuality and nudity is relatively straightforward. Ringrow (2016) describes sensuality as transcending nudity, aiming to evoke sexual feelings through a variety of sensory stimuli beyond the visual. Visual elements like parted lips have been associated with sexual arousal, yet cultural contexts are central to how these cues are rendered socially. In some cultures, including Middle Eastern and South Asian cultures that feature in the TMRC, any form of female nudity in the public space is perceived as sexualized; in other cultures, however, nudity may be featured without overt sexualization. These distinctions may appear exceedingly subtle for social media propaganda, where the representation of women leans towards sensual tropes and veers towards objectification, with female avatars often showcasing their attractiveness by foregrounding their physical appeal (Davis, 2018; Rose et al., 2012).

The perception of sensuality also varies substantially across the groups and cultures targeted by influence operations in the TMRC database, chiefly because sensuality encompasses the myriad ways humans experience physical pleasure. Skin exposure and skin color, for instance, have played significant roles in shaping perceived notions of sensuality across social groups. Before skin color became primarily associated with race and class, skin tone carried sexual connotations (Frost, 1990). Sensuality is not only rooted in tangible biological factors; it is culturally and historically defined, with the Japanese manga-style portrayal of beauty marked by slender and elongated legs having shaped the more recent perceptions of sensuality in East Asian culture (Starr et al., 2020). Sensuality is nonetheless successfully commodified by the advertising and cosmetics industry to evoke sensory experiences that shape emotional engagement and brand loyalty (Roberts, 2005). In this context, sensuality is used to market products to women without their narrative control (Wolf, 1991).

Given the above, we parameterize sensuality as a binary value (T/F) measured on visual elements such skin exposure, voluptuous hair, body pose, and facial expressions, including half-smiles and parted lips. Our approach draws on the insights of sensuality as extending beyond mere nudity to evoke sexual feelings through diverse sensory stimuli. Visual cues are however context-dependent and as such we sought to consider the cultural nuances highlighted by Ringrow (2016) by employing coders outside Europe and North America. Our coding scheme also takes into account what has been termed 'beauty pornography' (Wolf, 1991) in reference to subtle visual elements like parted lips that are linked to sensuality and suggest implicit sexualization even in the absence of overt nudity. Intercoder reliability for the manually coded data was performed across the three coders who independently assessed a random sample of images, with Krippendorff's alpha of .739 indicating substantial agreement among coders despite the perennial challenges in defining sensuality.

Camera Angle

The angle of the camera is broadly defined by whether the subject is shot from above, below, or at eye level to frame the object through high, low, or neutral angle shots (Merkt et al., 2022). The perceptions linked to camera angles are influenced by evolutionary cues, social learning, and embodied cognition. Evolutionary cues stem from height signaling dominance and threat (Freedman, 1979), and social learning literature established that power associations with verticality begin in childhood and persist into adulthood (Schwartz et al., 1982). Research on embodied cognition, finally, posits that abstract concepts like power are based on physical experiences whereby higher positions translate to more power (IJzerman & Koole, 2011). Language reflects these perceptions by equating power with upward positions and lack of power with downward trajectories. It also associates vertical angles with dominance or subordination (Meyers-Levy & Peracchio, 1992).

Changes in camera angle can lead to significant and predictable changes in how the physical and personal characteristics of the photographed object are judged. Low-angle shots often make the object appear taller and stronger, and are thus employed to portray power and courage (Figure 2). Eye-level shots, on the other hand, connote a sense of equality, parity, or neutrality. High-angle shots tend to present the photographed subject as weaker or frail and manufacture a sense of vulnerability (Kraft, 1987). Notably, men tend to be depicted from low angles to suggest dominance and power, whereas women are often shown from high angles to suggest fragility or lower status. Memory recall for these traits was found to be more accurate than for the specific camera angles that influenced these judgments. Huang et al. (2002) also found that the perceived differences in height influenced the outcomes of group decision-making tasks over video chat, with taller participants having more influence over the group.

Type of Shot

The type of shot in visual storytelling allows for conveying emotions, emphasizing context, and connecting with viewers on a personal level by changing the camera's distance from the subject in focus. A common directing technique, camera distance is used to change the emphasis between the subject and the surrounding scene and it affects the audience's emotional involvement and identification with the photographed subject (Canini et al., 2011). These effects are explained by proxemics thresholds where the perceived closeness increases persuasiveness and likability to suggest a subliminal familiarity (Grayson & Coventry, 1998; Mehrabian & Williams, 1969; Patterson, 1968). The distance between the photographed subject and the viewer decreases as we move from long to mid-shots, close-ups, and big close-ups (Arijon, 1991; Hall, 1990), with the

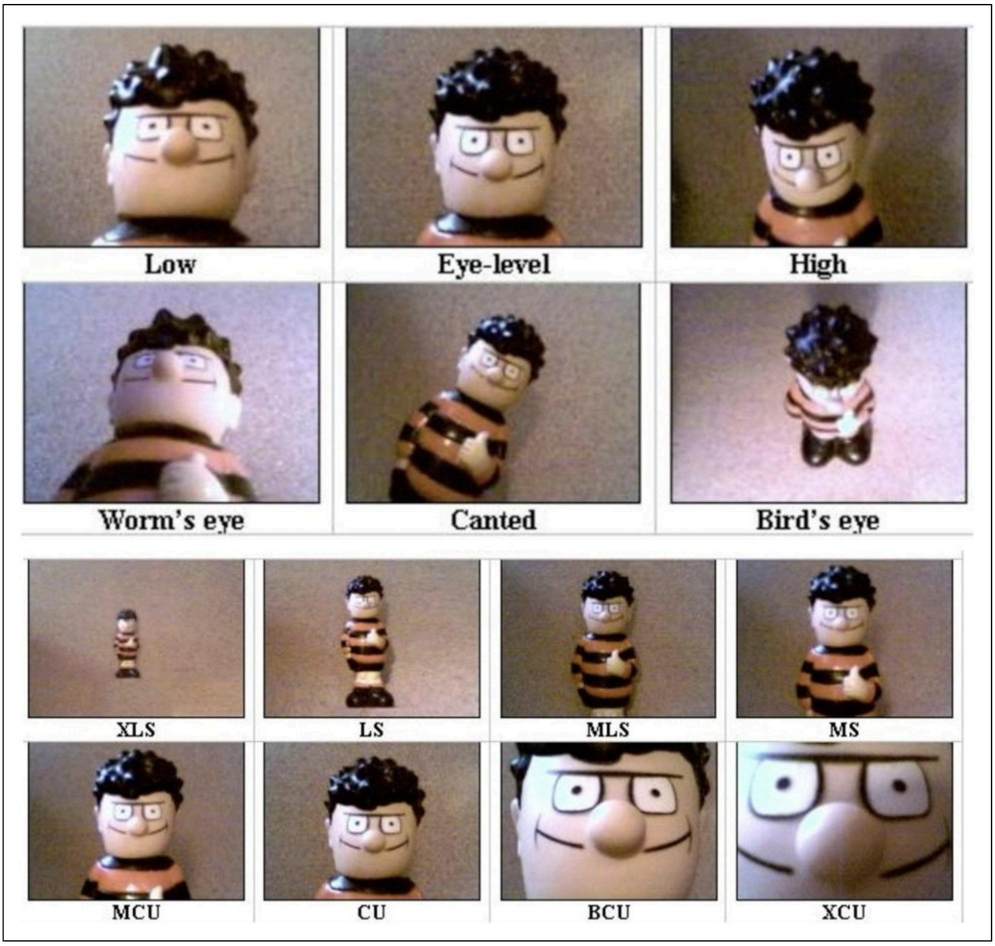

Figure 2. Camera angle and camera distance in visual composition (Chandler, 2001).

camera distance often dictating audience attention and closer shots garnering more focus (Wang & Cheong, 2009).

Long shots establish context by offering a wide perspective and including the subject and its environment, whereas medium shots cut at the waist, blending the subject with their setting to foreground a personal view of gestures and expressions that strike a balance between subjects and their surroundings. This is in sharp contrast to close-ups that fill the screen with the subject's face and shoulders for a detailed, intimate view of their expressions, and that are designed to highlight subjective experiences and evoke emotions (Canini et al., 2011). Large faces in photos also correlate positively with engagement, as larger objects attract more attention (Peng, 2021). The face-ism index, which measures face prominence in photographs, suggests that more visible faces are associated with perceptions of ambition and intelligence, whereas greater body prominence foregrounds traits like attractiveness and emotion (Archer et al., 1983). Figure 2 shows variations in camera angle and distance that are central to visual composition.

The model outputs both the predicted class and the confidence score for each prediction, with the probability distribution expressed in percentages across each dimension set against a threshold value to convert the continuous probabilities into discrete classes. A threshold of 75% is employed for binary classifications such as sensuality and gender, so if the predicted probability for a class is above this threshold we consider the outcome to be positive. A threshold of 50% is employed for multidimensional classes like race, type of shot, and camera angle so that the class with the highest probability (above 50%) is selected. In practical terms, the DeepFace model classifies gender and race based on the highest confidence scores for each category as described in the literature (Agarwal, 2018; Raj et al., 2020; Serengil & Ozpinar, 2020; Shorten & Khoshgoftaar, 2019). In the remainder of this study we evaluate the performance of this framework for the identification of fake social media profiles, with the codebase openly available on the project's repository.

TMRC

Twitter has curated a large database of influence operations that includes user accounts and the content posted by this cohort of users. Originally made available to researchers through their Civic and Election Integrity, this initiative evolved into a large database known as the Twitter Moderation Research Consortium. In addition to comprehensive visual information, the database offers information about the number of removed accounts, number of tweets, languages, hashtags, reported locations, and technical indicators of location for a broad spectrum of accounts from various regions around the world (Gadde & Roth, 2020; Roth, 2019). Since Twitter's acquisition by Elon Musk, however, data related to the Twitter Moderation Research Consortium is no longer available to researchers.

The TMRC is separated by influence operation, with specific datasets reflecting the geographic coverage of the campaign. The TMRC14_APAC_3, for instance, contains content from accounts that predominantly posted in Urdu. It details the removal of 568 accounts that collectively posted over 4 million tweets reportedly from Pakistan, often featuring sensual women (see Figure 3). The geographic coverage of the database has a considerable impact on the analysis of the data, particularly for context and culturally dependent variables like sensuality. To this end, we manually coded images with characteristics associated with femininity, corporeality, and eroticism across the many regional contexts and the varied perceptions of sensuality that emerged from the data. The selection ensured that the training corpus included a balanced representation of sensual images across different cultures while also preventing misclassification of non-sensual images that may share similar visual properties with sensual ones. A total of 2000 images were manually coded across each category to establish a benchmark. For camera angle, we also relied on the CAD2 training set to address variability in more controlled conditions.

Results

Sensuality

We start by identifying sensual images in the entire TMRC database and inspecting the performance of models generated by Teachable Machine against Google Vision API (Figure 3). Benchmarking for this comparative analysis was set using a hand-curated dataset of 500 images for each category from the TMRC dataset, distinctly labeled as sensual or non-sensual. Key performance indicators include an accuracy of .87 and .72 and a precision of .65 and .35 for Teachable Machine and Vision API, respectively. They also include a recall of .48 and .80, an F1 of .55 and .49, and a specificity of .94 and .71 for Teachable Machine and Vision API, respectively. Figure 4 shows the confusion matrices of both models. We also inspected the Receiver Operating Characteristic (ROC) curve for further performance comparison between Teachable Machine and Vision API (Figure 4).

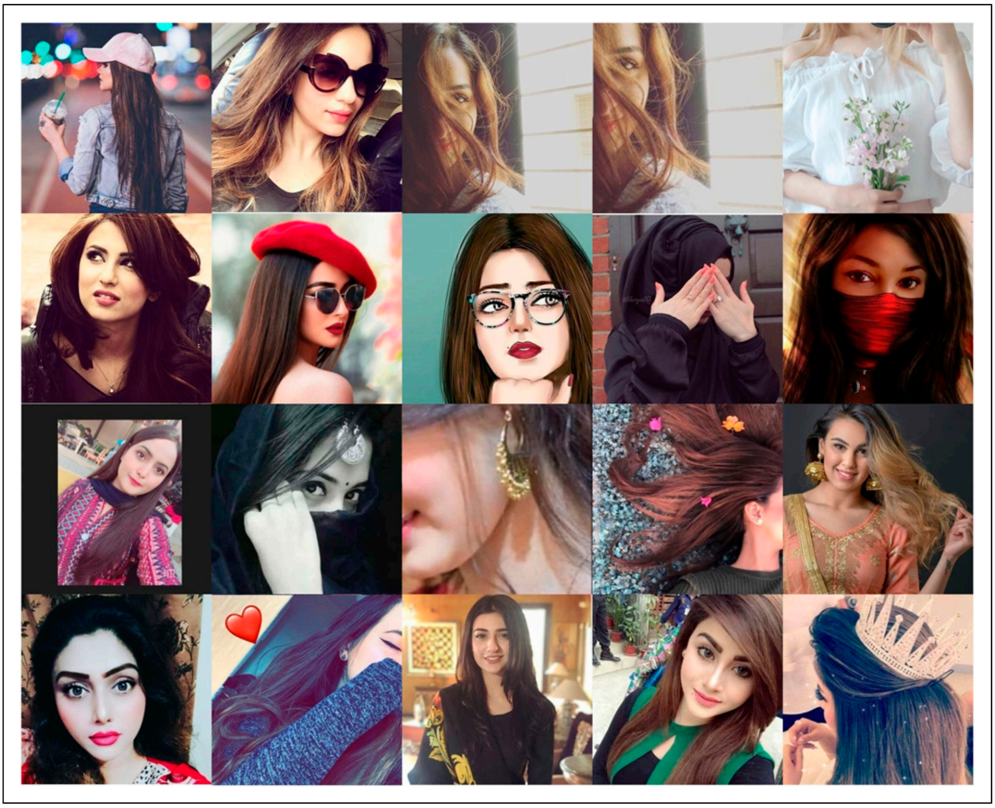

Figure 3. Sample profiles from TMRC14_APAC_3 identified as sensual.

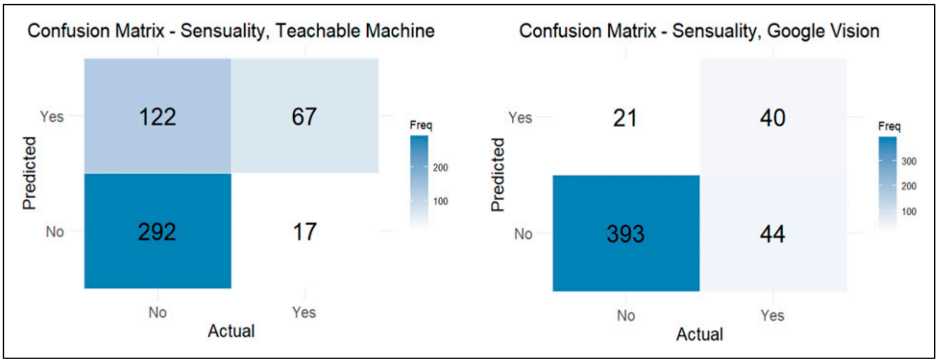

Figure 4. Confusion matrices for the classification of sensual images on Teachable Machine (a) and Vision API (b).

While Teachable Machine presents a notably higher recall rate of .80, precision was significantly lower at .35 compared to Vision API's score of .65. Vision API also showed higher accuracy (.87) compared to the Teachable Machine (.72). This suggests that while Teachable Machine was better at correctly identifying sensual images, it also misclassified a larger number of non-sensual images as sensual. The F1-Score, which balances precision and recall, is slightly higher for Vision API (.55) compared with Teachable Machine (.49), and in specificity Vision API outperforms Teachable Machine with a score of .94 against .71, indicating its superior ability to correctly classify non-sensual images. But the Receiver Operating Characteristic (ROC) curve area provides some nuance to these findings. Indeed, the area under the curve (AUC) for Teachable Machine was .75, slightly exceeding Vision API's AUC of .71. In other words, while Vision API reported a higher overall accuracy, Teachable Machine showed a marginally better discriminative ability when considering the recall. Figure 5 unpacks these results.

Camera Angle

Given that the data entail categorical variables that have an inherent order (Low < Neutral < High), Kendall's tau-b (τb) correlation coefficient was chosen to assess the concordance between the predicted and actual values for camera angle. CAD2-MN displayed the weakest association with Kendall's tau-b of .0682. This was followed by CAD2-PN, which showed a slight positive concordance at Kendall's tau-b of .1196. CAD1-PN showed a mild agreement with tau-b of .1902, while CAD1-MN stood out with a more robust moderate concordance of Kendall's tau-b of .3044. These values highlight the gradation in the models' effectiveness in maintaining the inherent order of the camera angle categories. The increasing positive values show that there are more concordant pairs where the order is maintained compared with discordant pairs.

The models trained on CAD1 outperformed those trained on CAD2. For the same data, the models trained on MobileNet performed slightly better than those trained on PoseNet. We further inspected their performance for each category and found that all models offered high precision for neutral angles, but the recall for neutral images is higher for CAD1-PN and lower in comparison with CAD1-MN. We observe that the poorer-performing models CAD2-MN and CAD2-PN

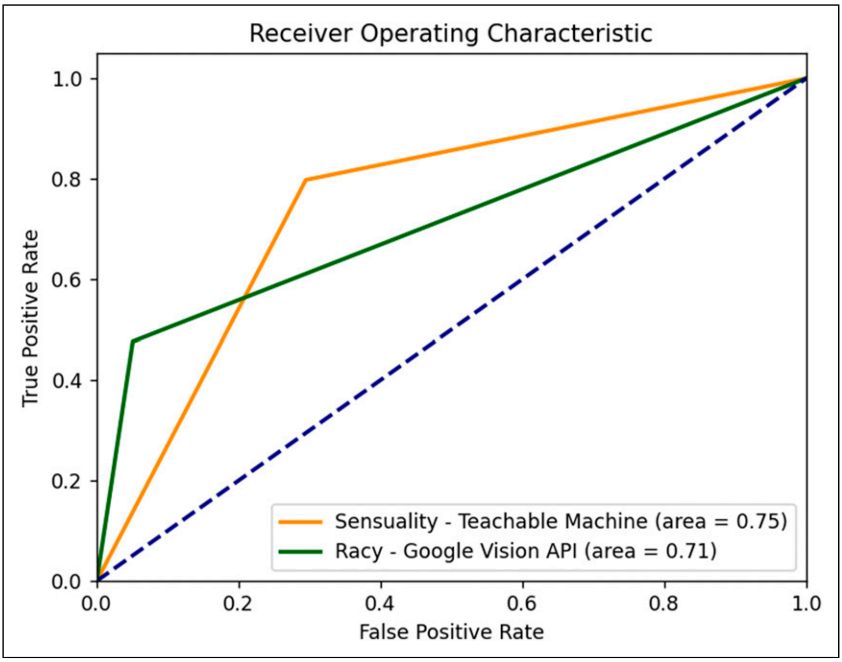

Figure 5. Receiver Operating Characteristic (ROC) curves for Google's Teachable Machine and Vision API in identifying sensual images in the TMRC dataset.

exhibit higher recall values for both high and low angles, but their precision values are notably lower. The confusion matrices show a significant overlap between neutral and other categories for the low-performing models (Figure 6). Upon inspecting the ROC curves for the better-performing model, we infer that CAD1-MN does a better job of distinguishing between categories shown by its slightly higher AUC values (Figure 7).

We relied on a manually classified dataset (n = 620) where all six categories were equally represented to estimate model performance for camera angle and type of shot. As this model includes categorical, non-ordinal variables, Cramer's V was used to quantify the association between the predicted and actual classifications. Both models performed well with Cramer's V values of .756 and .773 for TSD-MN and TSD-PN models, respectively. The higher Cramer's V value translates to higher precision, accuracy, recall, and F1-scores overall, particularly for big close-ups and long shots. Illustrations were included only under TSD-MN and are classified with a high precision score of .95. The performance of TSD-PN in mid-shot is superior to that of TSD-MN, but both models perform poorly for close-ups and selfies, with the confusion matrices showing that the models fail to accurately distinguish between selfies and close-ups. To a lesser extent, this is also the case for mid-shots (waist shots) and long-shots classified by TSD-MN. Looking at the ROC curves, however, we find that TSD-MN has overall more differentiating capabilities than TSD-PN as indicated by the higher AUC values (Figure 7).

Gender and Race

Identifying gender and race from profile images is a relatively established but fraught area in computer visual analysis (Buolamwini & Gebru, 2018). We manually classified the perceived gender and race displayed in a subset of the profile images and tested the accuracy of the DeepFace algorithm. The algorithm performed particularly well in detecting the gender dimension, with a precision of 99.5% and a recall of 99.5% for men, though precision and recall for women were lower at 73.8% for both parameters. Taken together, the DeepFace algorithm achieved an accuracy rate of 84.44% in gender identification, with an F1-score of 84.43%. Accuracy for racial categories varied across subgroups. The model correctly identified Asians 72 times, but often misclassified Asians as Latino, Middle Eastern, or White.

Overall, the model was moderately effective by achieving an accuracy of 62.42%, precision of 63.49%, and recall of 62.42%. The F1-score, which balances precision and recall, was slightly

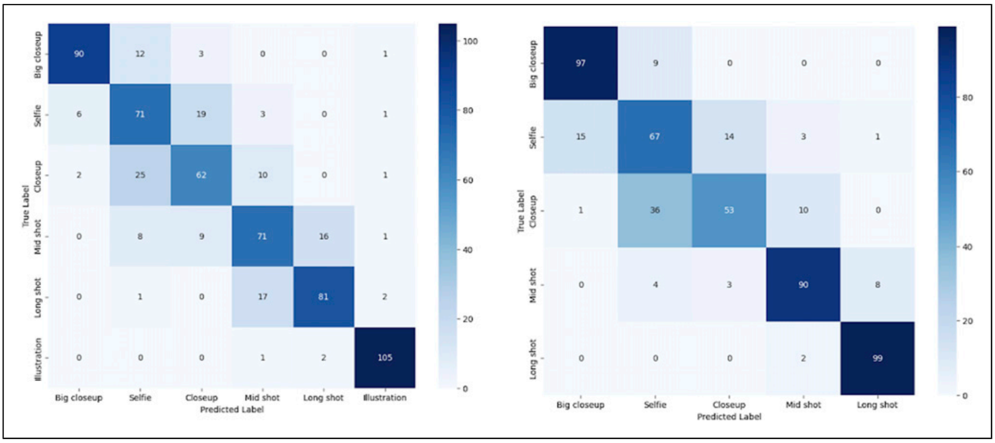

Figure 6. Confusion matrices for camera angle models.

lower at 60.39% compared with the gender classifier. Figure 8 shows the confusion matrices for the categories gender and race. The combined classifier yielded a high accuracy rate in gender classification for the Latino Hispanic group, with men being identified correctly 56 times and women 26 times, although 10 women were misclassified as men. This translates to an accuracy rate of 89.13% and an F1-score of 87.84%. The ROC curve features an impressive area of .86, indicating strong diagnostic capability, with a precision of .9077 and recall of .8913, another marker of a high rate of correct positive predictions.

But the intersection between gender and race showed greater variability. The algorithm was good at identifying Asians, with a perfect record for men and a high success rate for Asian women. Gender accuracy for Asians was 80.77%, with an F1-score of 80.38%. The ROC curve for Asians is .86, with a precision of .8780 and a recall of .8076, comparable to that of the Latino Hispanic subgroup and indicating solid model performance. The algorithm was also effective at identifying gender within the Indian subgroup, with an accuracy of 89.33%, a precision of .9131, and a recall of .8933. The F1-score was similarly high at 89.31% with the ROC at .90.

The algorithm's performance for the Middle Eastern subgroup was also effective, with accuracy at 88.89%, precision of .9099, and recall of .8888. The F1-score of 88.84% and the ROC

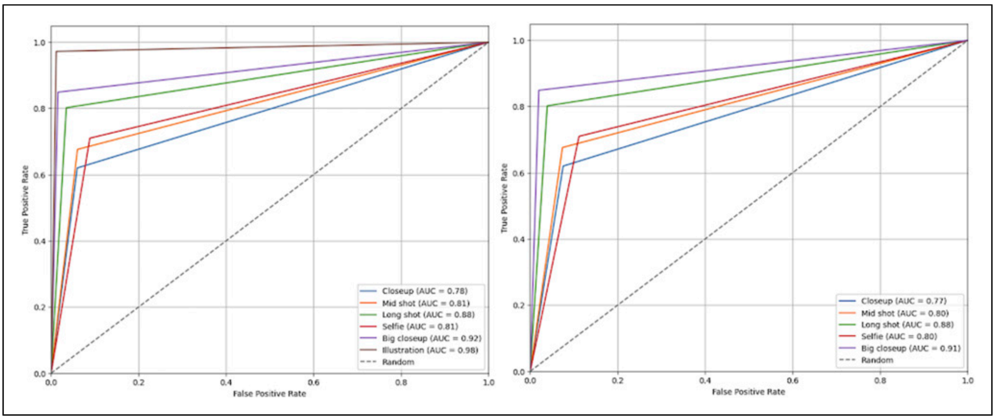

Figure 7. ROC curves for camera angle models.

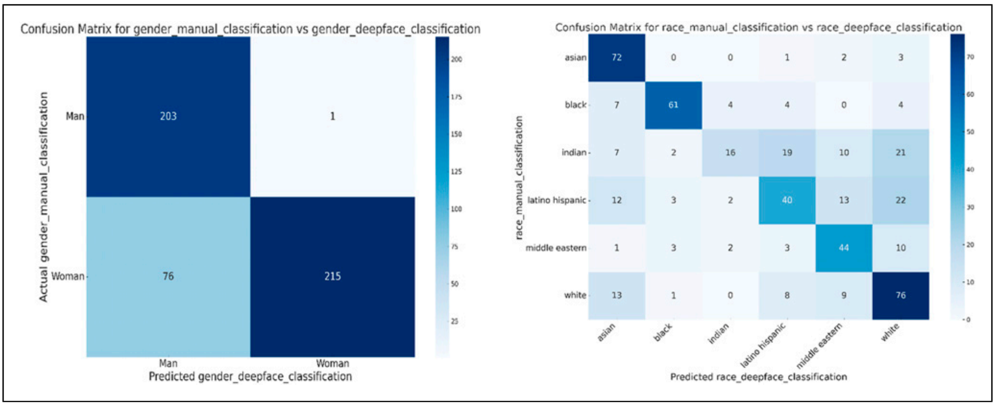

Figure 8. (a) Confusion matrix for gender (manual vs. DeepFace classification). (b) Confusion matrix for race (manual vs. DeepFace classification).

curve of .89 are similarly indicators of good performance. The least accurate subgroup was of individuals of African descent. While the algorithm could perfectly identify men, it had only 50% accuracy for women, leading to an overall accuracy rate of 67.5% and an F1-score of 67.48%. The ROC curve was also lower at .75, with precision at .8314 and recall at .675, a result of the model's moderate ability to detect true positive cases. Finally, the algorithm was also effective at the intersection of gender and race for Caucasians, with white men and women identified correctly at rates of 96.6% and 87.2%, respectively, which is in line with the broader biases identified in mainstream machine learning algorithms (Buolamwini & Gebru, 2018; Kleinberg et al., 2018). The model's accuracy stood at 89.72% with an F1-score of 88.05% and an ROC of .92 AUC. Precision was similarly high at .9181, indicating accurate positive predictions, with strong recall at .8971, reflecting the model's capabilities in identifying true positives. Figure 9 shows the confusion matrices for the intersection of gender and race across subgroups.

Discussion

The comparison between Google's Vision API and Teachable Machine in identifying sensual images provides insightful revelations about the performance of both models. The model built with Teachable Machine to identify sensual images in profile pictures of social media propaganda underperformed relative to the Vision API 'racy' category, a classifier that identifies content that is suggestive or provocative, often characterized by skimpy attire, provocative poses, or close-ups of sensitive body areas. Unsurprisingly, Vision API shows superior performance in terms of accuracy, precision, and specificity, likely due to being trained on ImageNet, a repository of approximately 100 million images. This vast exposure allows the model to generalize better across a myriad of image types and contexts, leading to classifications with higher accuracy and precision. In contrast, Teachable Machine may face constraints due to the specificity of its training data even if it leverages MobileNet's transfer learning, which was originally designed to discern between 1000 classes and may therefore limit its ability to generalize across contexts.

This limitation could be a significant contributor to Teachable Machine's relatively lower accuracy and precision. While it offers the benefits of customization, it may lack the sophisticated fine-tuning and continuous optimization dedicated to products like Vision API. Despite being outperformed in accuracy, precision, and specificity, the Teachable Machine algorithm surpasses Vision API in recall, achieving a score of .80 against Vision API's .48. Furthermore, the ROC curve for the Teachable Machine algorithm stands at .75, marginally outperforming Vision API's score of .71. The F-scores of both models are also comparable, with Teachable Machine scoring .49 against Vision API's score of .55. Perhaps more interestingly, the false positives triggered by Teachable Machine when classifying sensual images follow a discernible pattern, with close-up shots of white or fair-skinned women under bright lighting conditions having been often mistakenly classified as sensual. Interestingly, this pattern observed in the misclassification speaks to long-running social constructions around sensuality, with Frost (1990) arguing that skin color and tone initially held sexual connotations before becoming associated with race and class.

The identification of camera angles and type of shots is greatly dependent on the quality of the facial images available. Clear and sharp images yield much more accurate verification and recognition (Agarwal, 2018), but many of the images in the TMRC database are of subpar quality. The DeepFace algorithm may further worsen the signal by focusing on and cropping facial regions, which leads to further quality loss. Proactive image preprocessing can mitigate some of the effects associated with cropping and improve classifier accuracy. By fortifying the quality of images prior to analysis, the performance of facial recognition systems can be substantially enhanced (Raj et al., 2020). Additionally, the abundance of AI beautification filters built into

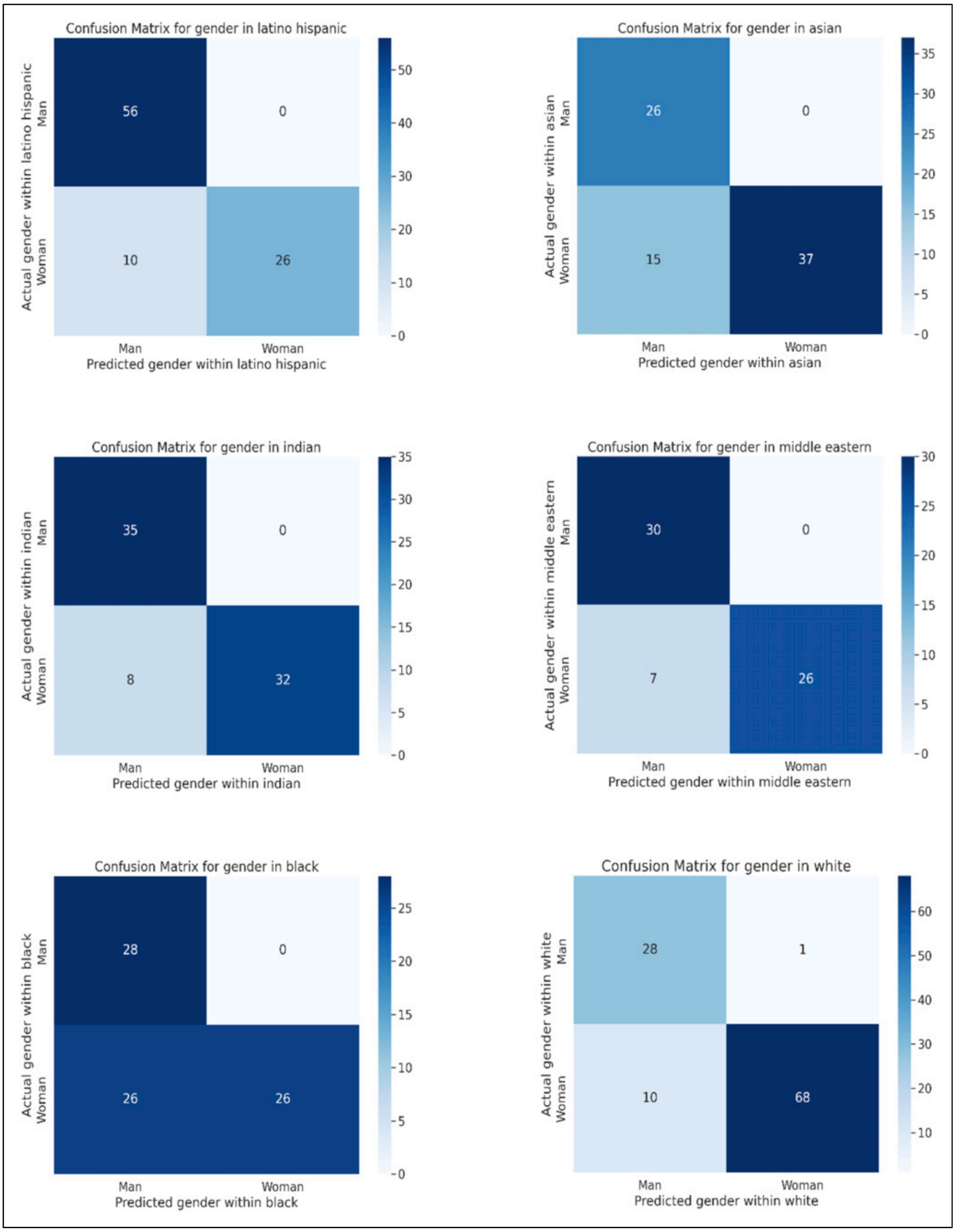

Figure 9. Confusion matrices for the intersection of gender and the following racial categories: Latino Hispanic subgroup (a); Asians (b); Indians (c); Middle Easterners (d); Blacks (e); and Caucasians (f).

smartphones can also distort facial landmarks that are critical for algorithms that identify photo composition, but also categories like gender and race (Broz, 2022; Yang, 2021). Such alterations by AI filters challenge facial recognition systems like DeepFace and lead to potential misclassification when features fail to align with the trained data.

This seems to be the case in the observed performances of models CAD1 and CAD2. While the latter was trained on a larger dataset, it was overperformed by CAD1 likely due to the stratified random sampling of TMRC folders and the resulting overfitting of CAD2. Similarly, MobileNet overperformed PoseNet models, which rely on landmarks for image composition. This is likely a result of occlusion, a significant hindrance in image processing that diminishes available visual data (Antoniadis, 2022). In PoseNet, occlusion can lead to inaccurate detection of key points essential for angle estimation, and is particularly problematic in common image scenarios like selfies. In contrast, models analyzing shot types showed better performance, indicating effective feature extraction and limited occlusion, as the focus was on the screen space occupied by the subject. PoseNet models nonetheless struggled with big close-ups and selfies due to their reliance on body landmark key points, which are compromised by shot proximity and body orientation.

Finally, gender classification is an inherently challenging task due to the complex, non-binary dimension of gender identity, in sharp contrast to more stable characteristics tied to sexual dimorphism (Butler, 1999; West & Zimmerman, 1987). As gender identity is largely fluid and constructed, and ultimately influenced by intersectionality with race, class, and age (Cole, 2009; Crenshaw, 1997), it presents a formidable challenge for tools like the DeepFace, which often struggles with gender classification due to the dynamic nature of such identities, particularly for visuals that do not conform to the traditional gender binary or whose appearance has been altered by camera filters. This is in line with previous research that cautioned against gender classification based on facial structures without taking into account the contextual and social aspects of one's identity (Bastos et al., 2023). A more comprehensive model would analyze not just facial features, but also user-generated content like bios and self-reported names to portray the array of complex elements shaping one's gender identity on social platforms.

Conclusion

The framework detailed in this study can be leveraged to identify influence operation at scale given their reliance on stable visual markers such as sensuality and camera angle. Future research should expand the framework described in this study by incorporating additional conceptual dimensions beyond gender, race, sensuality, camera angle, and type of shot that are salient in political propaganda and the cottage industry of romance scams. This framework for computer vision communication should prove particularly relevant for social media research on Instagram, TikTok, and other services that rely on rich social media content. Indeed, the Keras model detailed in this study can be promptly applied to any set of images collected from social media platforms. To this end, it should be possible to set up a data collection pipeline based on hashtags and/or keywords that would then be processed to identify higher-than-average levels of sensual content that intersect with gender, race, emotion, and specific photographic frames. Such an automated approach could yield significant results in early-detection systems or be integrated into downstream visual social media research analysis.

Another immediate application of this framework, beyond the identification of inauthentic political campaigns, refers to the automated detection of fake social media profiles explored by the cottage industry of frauds and scams using desirable online personas. Leaked manuals of pig butchering scams explicitly mention that fake social media profiles of men should craft personas with developmental disorders or previous psychological trauma to exploit the victim's maternal love. Fake social media profiles of women, on the other hand, require no such backstory but include detailed information about their visual composition. These profiles should not be crude; they should feature a naughty but cute nickname and present attractive young females who appear wealthy and educated (Faux, 2023; Reddit, 2021). These guidelines are supposed to arouse the victim's inner dreams and goals and entail another dimension of social infiltration.

The framework could also be expanded to other directions of research by training visual characteristics not covered in our Keras model (e.g., the incidence of political slogans, military fatigues, or tactical gear). Further training of the data through Teachable Machine is likely to be required, but it would suit research agenda contending with information that needs to be processed efficiently and with relative ease of implementation. The most salient challenges in expanding the framework detailed in this study include the selection of an appropriate pretrained model that resonates with their data in terms of subjects, sources, and styles. To this end, an examination of the labels in the pretrained dataset can offer important insights into the potential alignment with the research objectives of the study, even if the availability of pretrained models remains limited (Zhang & Peng, 2022).

Ultimately, DeepFace and Teachable Machine offer a user-friendly machine learning framework that can effectively identify fake social media profiles, visual disinformation, and inauthentic social media activity by leveraging deep learning. While traditional machine learning frameworks employ systematic hyperparameter tuning, often using techniques like grid search or random search (Bergstra & Bengio, 2012), Teachable Machine relies on a basic trial-and-error method for setting the parameters. Future research seeking to expand on our model should begin by collecting visual data from social media that are then preprocessed through resizing, scaling, and data augmentation. The pre-trained models detailed in this study can then be used for feature extraction and fine-tuning on this dataset through transfer learning. The training involves building and customizing convolutional neural networks (CNNs) to identify visual markers typical of inauthentic and manipulated content. The final model, evaluated for accuracy and robustness, can ultimately be deployed to monitor and flag potentially deceptive visual content online in real time.

Teachable Machine, however, has limited interpretability tools, which are central for building trustworthy and transparent machine learning models (Doshi-Velez & Kim, 2017), and lacks support for data augmentation techniques that are important for image classification tasks, including random rotations, zooming, and flipping that artificially expand and diversify the training dataset (Shorten & Khoshgoftaar, 2019; Jung & Oh, 2021). The absence of such features is particularly salient in tasks that demand fine distinctions between categories and small-sized datasets, or where augmentation could significantly enhance model performance. These models, however, offer a tangible and effective framework for analyzing inauthentic social media campaigns at scale, with the Keras model trained for this study offering an off-the-shelf tool for research on influence operations exploiting gender, sensuality, and race to infiltrate target groups with messages that are segmented into groups with clear visual identities emphasizing selfies and sensual young women. Indeed, sensuality is quickly becoming a key variable associated with influence operations replicating the Kremlin-linked Internet Research Agency campaign, but future research should continue to expand the visual markers driving social media disinformation.

Appendix

Coding Variables

1. User
(Stated) user identity (string)
2. File name
Indexed file name for user profile (numeric)
3. Content description
300-characters field describing the image (string)
4. Racial category

Racial category of individuals featured in the image according to https://data.london.gov.uk/dataset/ethnic-groups-borough (Last updated Sep 2018). Includes explicit reference to ethnic minorities, such as Muslims in Europe and African-Americans in the U.S. Values: White, Black, Asian, Mixed, or Others (categorical)

5. Type of photo (frame)

Type of photo: super close up (selfie), big close up, close up, illustration, long shot, or medium shot (categorical)

6. Camera angle

Low angle, neutral, or high angle (categorical)

7. Photo quality

Professional versus amateur (binary)

8. Indoor/outdoor

Indoor versus outdoor (binary)

9. Number of people

Number of individuals in the photo (numerical)

10. Male

True or False (logical)

11. Female

True or False (logical)

12. Sensual

True or False (logical).

13. Emotion

(Perceived) user emotions in the photo: anxious, ashamed, bored, happy, ironic, sad, scared, serious, joyful, or tense (categorical)

14. Gender

Male, Female, Other, or NA (categorical)

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904, research grant #270353197 “The Visual Frames of Social Media Propaganda,” and Twitter, Inc. research grant 50069SS “The Brexit Value Space and the Geography of Online Echo Chambers.” The corresponding author also acknowledges support from the University College Dublin and OBRSS scheme (grants R21650 and R20825) and the National Council for Scientific and Technological Development (grant 406504/2022-9).

ORCID iD

Marco Bastos https://orcid.org/0000-0003-0480-1078

References

Agarwal, V. (2018). Deep face quality assessment (arXiv:1811.04346). arXiv. https://doi.org/10.48550/arXiv.1811.04346

Antoniadis, P. (2022, February 19). Image processing: Occlusions | Baeldung on computer science. https://www.baeldung.com/cs/image-processing-occlusions

Archer, D., Iritani, B., Kimes, D. D., & Barrios, M. (1983). Face-ism: Five studies of sex differences in facial prominence. *Journal of Personality and Social Psychology*, *45*(4), 725–735. https://doi.org/10.1037/0022-3514.45.4.725

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27. https://doi.org/10.1145/3274289

Arijon, D. (1991). *Grammar of the film language*. Silman-James Press. https://www.amazon.co.uk/Grammar-Film-Language-Daniel-Arijon/dp/187950507X

Bastos, M. T., & Farkas, J. (2019). “Donald Trump is my president!”: The Internet research agency propaganda machine. *Social Media + Society*, *5*(3), Article 205630511986546. https://doi.org/10.1177/2056305119865466

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Bastos, M. T., Mercea, D., & Goveia, F. (2023). Guy next door and implausibly attractive young women: The visual frames of social media propaganda. *New Media & Society*, *25*(8), 2014–2033. https://doi.org/10.1177/14614448211026580

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Bonilla-Silva, E. (1997). Rethinking racism: Toward a structural interpretation. *American Sociological Review*, *62*(3), 465–480. https://doi.org/10.2307/2657316

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (Mis)Representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics*, *26*(1), 277–299. https://doi.org/10.1177/1940161220964780

Broz, M. (2022, February 17). How many photos are there? (Statistics & trends in 2024). https://photutorial.com/photos-statistics/

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, *81*, 77–91.

Butler, J. (1999). *Gender trouble* (2nd ed.). Routledge. https://doi.org/10.4324/9780203902752

Canini, L., Benini, S., & Leonardi, R. (2011). Affective analysis on patterns of shot types in movies. In 2011 7th international symposium on image and signal processing and analysis (ISPA), Dubrovnik, Croatia, 04–06 September 2011, pp. 253–258.

Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffith, J., Jongejan, J., Pitaru, A., & Chen, A. (2020). Teachable machine: Approachable web-based tool for exploring machine learning classification. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, pp. 1–8. https://doi.org/10.1145/3334480.3382839

Çeliktutan, O., Ulukaya, S., & Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, *2013*(1), 13. https://doi.org/10.1186/1687-5281-2013-13

Chandler, D. (2001). The “grammar” of television and film. https://visual-memory.co.uk/daniel/Documents/short/gramtv.html

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*(3), 170–180. https://doi.org/10.1037/a0014564

Crenshaw, K. (1997). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In K. Maschke (Ed.), *Feminist legal theories* (p. 328). Routledge.

Davis, S. E. (2018). Objectification, sexualization, and misrepresentation: Social media and the college experience. *Social Media + Society*, *4*(3), Article 205630511878672. https://doi.org/10.1177/2056305118786727

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608v2

Faux, Z. (2023). *Number go up: Inside crypto's wild rise and staggering fall*. Penguin Random House.

Freedman, D. G. (1979). *Human sociobiology: A holistic approach*. Free Press. https://repository.library.georgetown.edu/handle/10822/548140

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, *40*(3), 560. https://doi.org/10.1177/0894439320914853

Frost, P. (1990). Fair women, dark men: The forgotten roots of colour prejudice. *History of European Ideas*, *12*(5), 669–679. https://doi.org/10.1016/0191-6599(90)90178-H

Gadde, V., & Roth, Y. (2018, October 17). Enabling further research of information operations on Twitter.

Gadde, V., & Roth, Y. (2020, November 12). An update on our work around the 2020 US Elections. https://web.archive.org/web/20210612062339/https://blog.twitter.com/en_us/topics/company/2020/2020-election-update

Garimella, K., & Eckles, D. (2020). Images and misinformation in political groups: Evidence from WhatsApp in India (arXiv:2005.09784). arXiv. https://arxiv.org/abs/2005.09784

Google. (2017). Teachable machine. https://web.archive.org/web/20171003160620/https://teachablemachine.withgoogle.com

Grayson, D., & Coventry, L. (1998). The effects of visual proxemic information in video mediated communication. *ACM SIGCHI Bulletin*, *30*(3), 30–39. https://doi.org/10.1145/565711.565713

Hall, E. T. (1990). *The hidden dimension, reprint*. Anchor Books.

Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

Harvey, D., & Roth, Y. (2018, October 1). An update on our elections integrity work. https://web.archive.org/web/20210624004136/https://blog.twitter.com/en_us/topics/company/2018/an-update-on-our-elections-integrity-work

Huang, W., Olson, J. S., & Olson, G. M. (2002). Camera angle affects dominance in video-mediated communication. In CHI '02 Extended Abstracts on Human Factors in Computing Systems, Minneapolis, MN, USA, pp. 716–717. https://doi.org/10.1145/506443.506562

Ijzerman, H., & Koole, S. L. (2011). From perceptual rags to metaphoric riches—bodily, social, and cultural constraints on sociocognitive metaphors: Comment on Landau, Meier, and Keefer (2010). *Psychological Bulletin*, *137*(2), 355–361. https://doi.org/10.1037/a0022373

Iyer, A., Webster, J., Hornsey, M. J., & Vanman, E. J. (2014). Understanding the power of the picture: The effect of image content on emotional and political responses to terrorism. *Journal of Applied Social Psychology*, *44*(7), 511–521. https://doi.org/10.1111/jasp.12243

Joo, J., & Steinert-Threlkeld, Z. C. (2018). Image as data: Automated visual content analysis for political science (arXiv:1810.01544). arXiv. https://arxiv.org/abs/1810.01544

Jung, H., & Oh, Y. (2021). Towards better explanations of class activation mapping (arXiv:2102.05228). arXiv. https://doi.org/10.48550/arXiv.2102.05228

Khatib, S. M. (1989). Race and credibility in persuasive communications. *Journal of Black Studies*, *19*(3), 361–373. https://doi.org/10.1177/002193478901900306

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/laz001

Kraft, R. N. (1987). The influence of camera angle on comprehension and retention of pictorial events. *Memory & Cognition*, *15*(4), 291–307. https://doi.org/10.3758/BF03197032

Lewontin, R. C. (1972). The apportionment of human diversity. In *Evolutionary biology* (pp. 381–398). Springer.

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research*, *57*(1), 1–19. https://doi.org/10.1177/0022243719881113

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. (2016). Analyzing personality through social media profile picture choice. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 1. https://doi.org/10.1609/icwsm.v10i1.14738

Lu, Y., & Pan, J. (2022). The pervasive presence of Chinese government content on douyin trending videos. *Computational Communication Research*, *4*(1), 68. https://doi.org/10.5117/CCR2022.2.002.LU

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Matz, S. C., Segalin, C., Stillwell, D., Müller, S. R., & Bos, M. W. (2019). Predicting the personal appeal of marketing images using computational methods. *Journal of Consumer Psychology*, *29*(3), 370–390. https://doi.org/10.1002/jcpy.1092

Mehrabian, A., & Williams, M. (1969). Nonverval concomitants of perceived and intended persuasiveness. *Journal of Personality and Social Psychology*, *13*(1), 37–58. https://doi.org/10.1037/h0027993

Merkt, M., Weingärtner, A.-L., & Schwan, S. (2022). Digital images are hard to resist: Teaching viewers about the effects of camera angle does not reduce the camera angle's impact on power judgments. *Acta Psychologica*, *229*, Article 103687. https://doi.org/10.1016/j.actpsy.2022.103687

Metaxas, P., & Mustafaraj, E. (2010). From obscurity to prominence in minutes: Political speech and real-time search. In Proceedings of the WebSci conference, Raleigh, NC, USA.

Meyers-Levy, J., & Peracchio, L. A. (1992). Getting an angle in advertising: The effect of camera angle on product evaluations. *Journal of Marketing Research*, *29*(4), 454–461. https://doi.org/10.2307/3172711

Muscanell, N. L., & Guadagno, R. E. (2012). Make new friends or keep the old: Gender and personality differences in social networking use. *Computers in Human Behavior*, *28*(1), 107–112. https://doi.org/10.1016/j.chb.2011.08.016

Omi, M., & Winant, H. (2014). *Racial formation in the United States*. Routledge.

Patterson, M. (1968). Spatial factors in social interactions. *Human Relations*, *21*(4), 351–361. https://doi.org/10.1177/001872676802100403

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, *68*(5), 920–941. https://doi.org/10.1093/joc/jqy041

Peng, Y. (2021). What makes politicians' Instagram posts popular? Analyzing social media strategies of candidates and office holders with computer vision. *The International Journal of Press/Politics*, *26*(1), 143–166. https://doi.org/10.1177/1940161220964769

Peng, Y., & Jemmott, J. B. III (2018). Feast for the eyes: Effects of food perceptions and computer vision features on food photo popularity. *International Journal of Communication*, *12*, 313–336.

Peng, Y., Lu, Y., & Shen, C. (2023). An agenda for studying credibility perceptions of visual misinformation. *Political Communication*, *40*(2), 225–237. https://doi.org/10.1080/10584609.2023.2175398

Powell, T. E., Boomgaarden, H. G., De Swert, K., & de Vreese, C. H. C. H. (2015). A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication*, *65*(6), 997–1017. https://doi.org/10.1111/jcom.12184

Qian, S., Shen, C., & Zhang, J. (2022). Fighting cheapfakes: Using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication*, *28*(1), Article zmac024. https://doi.org/10.1093/jcmc/zmac024

Raj, R. J. S., Shobana, S. J., Pustokhina, I. V., Pustokhin, D. A., Gupta, D., & Shankar, K. (2020). Optimal feature selection-based medical image classification using deep learning model in Internet of medical things. *IEEE Access*, *8*, 58006–58017. https://doi.org/10.1109/ACCESS.2020.2981337

Reddit. (2021). Pig-butchering scam/Sha Zhu Pan—photos of training manuals. https://www.reddit.com/r/Scams/comments/njimju/pigbutchering_scam_sha_zhu_pan_photos_of_training/?utm_source=share&utm_medium=web2x&context=3

Ringrow, H. (2016). *The Language of cosmetics advertising*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-55798-8

Roberts, K. (2005). *Lovemarks: The future beyond brands*. Powerhouse Books.

Rose, J., Mackey-Kallis, S., Shyles, L., Barry, K., Biagini, D., Hart, C., & Jack, L. (2012). Face it: The impact of gender on social media images. *Communication Quarterly*, *60*(5), 588–607. https://doi.org/10.1080/01463373.2012.725005

Roth, Y. (2019, January 31). Empowering further research of potential information operations. https://web.archive.org/web/20210613173536/https://blog.twitter.com/en_us/topics/company/2019/further_research_information_operations

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, *4*(1), 381–402. https://doi.org/10.1007/s42001-020-00084-7

Schwartz, B., Tesser, A., & Powell, E. (1982). Dominance cues in nonverbal behavior. *Social Psychology Quarterly*, *45*(2), 114–120. https://doi.org/10.2307/3033934

Seo, H. (2019). Visual propaganda and social media. In *The Sage handbook of propaganda*. Sage.

Serengil, S. I., & Ozpinar, A. (2020). LightFace: A hybrid deep face recognition framework. In 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, pp. 1–5. https://doi.org/10.1109/ASYU50717.2020.9259802

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Smedley, A., & Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American Psychologist*, *60*(1), 16–26. https://doi.org/10.1037/0003-066X.60.1.16

Starr, R. L., Wang, T., & Go, C. (2020). Sexuality vs. sensuality: The multimodal construction of affective stance in Chinese ASMR performances. *Journal of SocioLinguistics*, *24*(4), 492–513. https://doi.org/10.1111/josl.12410

Stencel, M., & Luther, J. (2021). *Fact-checking census shows slower growth*. Duke.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 12, pp. 1–9.

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

Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, *62*(1), 78–97. https://doi.org/10.1111/j.1460-2466.2011.01619.x

Torres, M., & Cantú, F. (2022). Learning to see: Convolutional neural networks for the analysis of social science data. *Political Analysis*, *30*(1), 113–131. https://doi.org/10.1017/pan.2021.9

Twitter. (2019). *Election integrity policy*. Twitter, Inc. https://web.archive.org/web/20190428071045/https://help.twitter.com/en/rules-and-policies/election-integrity-policy

Twitter. (2021). *Civic integrity*. Twitter, Inc. https://web.archive.org/web/20210130191423/https://about.twitter.com/en/our-priorities/civic-integrity

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1), Article 205630512090340. https://doi.org/10.1177/2056305120903408

von Sikorski, C. (2022). Visual polarisation: Examining the interplay of visual cues and media trust on the evaluation of political candidates. *Journalism*, *23*(9), 1900–1918. https://doi.org/10.1177/1464884920987680

Wang, H. L., & Cheong, L.-F. (2009). Taxonomy of directing semantics for film shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *19*(10), 1529–1542. https://doi.org/10.1109/TCSVT.2009.2022705

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society*, *1*(2), 125–151. https://doi.org/10.1177/0891243287001002002

Williams, N. W., Casas, A., & Wilkerson, J. D. (2020). *Images as data for social science research: An introduction to convolutional neural nets for image classification*. Cambridge University Press.

Wolf, N. (1991). *The beauty myth: How images of beauty are used against women*. Vintage.

Yang, Y. (2021). Smartphone photography and its socio-economic life in China: An ethnographic analysis. *Global Media and China*, *6*(3), 259–280. https://doi.org/10.1177/20594364211005058

Zhang, H., & Peng, Y. (2022). Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociological Methods & Research*, Onlinefirst. https://doi.org/10.1177/00491241221082603

Zhang, N., Luo, J., & Gao, W. (2020). Research on face detection technology based on MTCNN. In 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 25–27 September 2020, 154–158. https://doi.org/10.1109/ICCNEA50255.2020.00040

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. https://doi.org/10.1108/AJIM-09-2013-0083

Author Biographies

Noel George holds an MSc in Human–Computer Interaction from the University College Dublin and is a graduate from Mahatma Gandhi University, Kerala (B.Tech in Computer Science and Engineering).

Azhar Sham is a graduate of University College Dublin (MSc in Human-Computer Interaction, 2023) and the University of Kerala (B.Tech in Computer Science and Engineering, 2019). He specializes in UX, AI, and data analytics with a focus on human data interaction.

Thanvi Ajith is a B.Tech graduate from College of Engineering Trivandrum, with a Master's in Human Resource Management from Indian Institute of Technology, Kharagpur and a Master's in Human Computer Interaction from University College Dublin, Ireland.

Marco Bastos is the University College Dublin Ad Astra Fellow at the School of Information and Communication Studies and Senior Lecturer in Media and Communication in the Department of Media, Culture and Creative Industries at City, University of London. He is the author of *Spatializing social media: social networks online and offline* (Routledge, 2022) and *Brexit, tweeted: polarization and social media manipulation* (Bristol University Press, 2024). His research leverages computational methods and network science to explore the intersection of communication and critical data studies.