Essays on Visual Marketing

Evgeny Pavlov

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Natalie Mizik, Chair

Oliver Rutz

Hema Yoganarasimhan

Program Authorized to Offer Degree:

Foster School of Business

University of Washington

Abstract

Essays on Visual Marketing

Evgeny Pavlov

Chair of the Supervisory Committee:
J. Gary Shansby Endowed Chair in Marketing Strategy, Professor Natalie Mizik
Foster School of Business

As firms are embracing visual platforms in their marketing and branding efforts, little research exists on the relative effectiveness of visual versus text-based marketing efforts. In essays 1 and 2, we develop a quantitative framework to study how text and visual components of firm communications affect consumer engagement with firm-generated social media content. First, we quantify the emotional loading of text and imagery on sentiment and arousal/motivation-to-act dimensions. We use existing NLP tools for text and use machine learning and computer vision to develop and train sentiment and arousal classification models for imagery. We use four emotion modalities as predictors of visual emotion: (1) elements of design (low-level visual features) such as color, texture, shape, lines, curves, corners, edges, and orientation; (2) high-level visual objects/concepts (e.g., adventure, action, leisure, danger, etc.); (3) human facial expressions; and (4) text embedded in the image. We find that elements of design and high-level visual objects are the most important predictors of visual emotion. Our model achieves accuracy over 80%. Next, we apply the procedure to an empirical analysis of engagement (retweeting) with firm-generated content based on 1.3M tweets of 600+ brands from 11 categories, posted

since 2008. Our findings suggest that over the years, consumers have developed resistance to persuasion messaging using high motivation-to-act text. We do not find a similar decline in effectiveness for high motivation-to-act imagery. We find significant heterogeneity of image effects by industry, with positive and high motivation-to-act imagery being the most engaging for quick-service restaurants, and negative imagery being the most engaging for charities/non-profits. In essay 3, we study the effects of the face and gaze of models in the product images on outcomes such as product clicks, orders, and returns. We use deep-learning algorithms for face detection and gaze-following in the context of 57,088 apparel products from 22 categories on a large Chinese e-commerce website. We find that product images that include the model's face receive more clicks. However, higher prominence of the face leads to fewer clicks, but more ordering of the product. We also observe that the "direct" gaze and "downwards" gaze of the models in the images lead to fewer clicks and orders than "sideways" gazes. We offer potential explanations based on gaze psychology.

Table of contents

Acknowledgments

I would like to thank my advisor and mentor, Natalie Mizik. Not only is she a highly competent, patient, visionary, and optimistic academic advisor, but also always cares about me on a personal level. She encouraged me to do an industry internship (which is a rare and open-minded move for an academic advisor), counseled me extensively on the academic market, and has smart and forward-looking advice on broader aspects of life. I always weigh her opinion very highly and appreciate her perspective.

I am also very grateful to my committee members, Oliver Rutz and Hema Yoganarasimhan, for their advice, encouragement, support, and caring. Their feedback improved the research substantially. I thank Linda Shapiro for agreeing to be the GSR.

This PhD journey would not be possible without the full understanding, encouragement, and support of my family (Rita, Oleg, Anton, Lena, Yuri, Lida, Sasha, and Lyuba). They hold a major stake in whatever I am doing, because they dedicated so much care, love, and effort to raise and train me.

I am grateful to all the professors of the MIB department whose doctoral seminars I was lucky to attend, including Nidhi Agrawal, Lea Dunn, Natalie Mizik, Robert Palmatier, Oliver Rutz, Jeffrey Shulman, Richard Yalch, and Hema Yoganarasimhan. The seminars were critical in building breadth and depth of understanding of the multifaceted field of marketing, as well as in honing some essential technical skills.

I am also very thankful to all the professors who found time to listen to my research, give advice, and prepare me for the academic job market. Specifically, I thank my dissertation committee members and Abhishek Borah, Lea Dunn, Mark Forehand, Shailendra Jain, Simha

Mummalaneni, Amin Sayedi, Francesca Valsesia, and Scott Wallace. I thank Zhuping Liu for providing the data and advice for the essay 3.

What made my PhD time truly stimulating, fun, energetic, and heart-warming were my fellow PhD students. I thank Amir Fazli, Behnaz Bojd, and Aravinda Garimella for countless happy moments, Omid Rafieian and Shahryar Doosti for stimulating conversations, Chethana Achar for invaluable advice, and Melissa Rhee, Mohammad Arbabian, Maria Mitkina, Dmitry Brizhatyuk, Vladimir Dashkeev, Sareh Nabi, and Emisa Nategh for true friendship. Special gratitude goes to Ebrahim Barzegary, Pegah Jalali, Amin Zadkazemi, Vipul Aggarwal, Majid Majzoubi, Bita Hajihashemi, Mily Wang, David Shin, Katie Spangenberg, and Trevor Watkins. Every Foster PhD student/alum I know is an amazing and bright person, and I thank each of them for their energizing presence!

The doctoral program experience was made much smoother by the patient, caring, and very competent efforts of Jaime Banaag, Beau Kirkeby, Pam Tomaino, and Randy Kith. I am also thankful to Alex Kith, Amanda Meeks, and Christine Wainwright for helping me facilitate the research and teaching activities.

This journey would not be possible without the generosity of Dr. Wiley, Mr. Crowley, Scott Reynolds, and the Foster Doctoral Program. Thank you so much for providing extensive resources and funding to facilitate my research! I am also thankful to Mark Forehand and Shailendra Jain for being amazing MIB chairs who were always very supportive, helpful, and understanding when it came to funding and TA/RA assignment flexibility.

During my PhD time, I got to intern at Amazon Inc. as an economist. It was a very vibrant and growth-spurring period as I worked with extremely competent and inspiring experts. I thank

Patrick Bajari and Eric Zivot, as well as my managers and mentors Tim Graciano, Oleksiy Mnyshenko, Eduardo Jardim, and Laura Trucco.

Introduction

Firms are embracing visual platforms in their marketing and branding efforts. Thirty-four percent of 2016 annual marketing budgets were earmarked for creating, producing, and publishing visual content, up from 26% in 2014 (Koshy, 2016). According to industry surveys, 72% of marketers believe that visual marketing is more effective than text-based marketing (Gujral, 2015), and 90% of marketers would like to know the best ways to engage their audience with social media (Stelzner, 2016). However, little research has investigated the effectiveness of visual versus text-based marketing efforts in generating consumer engagement with the content.

Essays 1 and 2 study how text and visual components of firm communications influence consumer engagement with firm-generated social media content. The effects of text-based marketing have been addressed. Researchers have investigated text-based content in the form of online product reviews (Chevalier & Mayzlin, 2006; Timoshenko & Hauser, 2019), online chatter (Borah & Tellis, 2016; Tirunillai & Tellis, 2014), blogs (Mayzlin & Yoganarasimhan, 2012), and forums (Netzer et al., 2012). Various aspects of the text emotionality[1] of a message have been linked to engagement with the message.

Unlike text, research on visuals is nascent and sparse (e.g., Liu, Dzyabura, & Mizik, 2020; Jalali & Papatla, 2016). Existing studies on user engagement do not examine image effects beyond an image-presence indicator in a post (e.g., Stephen, Sciandra, & Inman, 2015). No large-scale research has systematically looked at *image characteristics* and linked them to consumer engagement. Also, researchers have not addressed the *relative effectiveness* of visual and text elements in generating engagement.

[1] Examples include emotional activation (Berger & Milkman, 2012), emotional content, banter, humor, and philanthropy (Lee, Hosanagar, & Nair, 2018), and outrageousness and humor (Tucker, 2014).

We focus on text and images and their interplay in influencing consumer engagement with firm-generated social media content. Specifically, we seek to answer the following research questions:

- Can we extract and quantify emotional content reflected in the text and visual components of corporate communications?
- Do the characteristics of text and image affect consumer engagement with firm-generated social media content, and, if yes, how?
- What is the relative contribution of text and imagery in generating engagement?
- Do the optimal text and image strategies differ across industries (e.g., charity vs. quick service restaurants)?

To answer these questions, we use some existing machine-learning and natural language processing (NLP) tools to extract the emotional loading of text. No such tools exist for images, and we develop and validate a machine-learning tool for extracting the emotional loading of an image. In our empirical application of the tool, we link the emotional loadings of text and images to consumer-engagement measures. We suggest managerial implications and recommendations for improving messaging (text and imagery) to increase consumer engagement.

In essay 1, we train and validate a tool that allows us to quantify the emotional loading of an image. All models of affect developed in the psychology literature (e.g., circumplex, PANA, vector models) emphasize *sentiment* and *arousal* (*activation/motivation-to-act*) as the two most important dimensions of human emotions. The tool that we propose hence scores images on dimensions of sentiment and arousal. We collect a novel ground-truth annotated set of images on the two primary dimensions of emotions: sentiment and arousal. In our study, 1,292 subjects annotated 13,386

original brand-generated images from Twitter. The images come from 637 brands from 11 categories.

Next, we extract visual characteristics from the images that correspond to four emotion modalities: (1) elements of design (low-level visual features) such as color, texture, shape, lines, curves, corners, edges, and orientation; (2) high-level visual objects/concepts (e.g., adventure, action, leisure, danger, etc.); (3) human facial expressions; and (4) text embedded in the image. We use gradient boosting to predict levels of image sentiment and arousal given the four-modality visual characteristics of this image. Our model achieves accuracy of over 80% and results in near-perfect predictions for images with extreme ground-truth levels of sentiment/arousal. The demo of the visual-emotion prediction tool is available at imagesentiment.com.

We also undertake a lab study to validate model-informed insights on the emotional impact of individual elements of design. For example, the model indicates color variety, non-smooth texture, and higher amounts of green and orange hues are associated with more positive sentiment. By contrast, many corners and red and pink hues are associated with high arousal. We validate these insights by modifying original firm-generated Tweet images to have less (or more) of a particular feature (e.g., decreased number of corners in the background of an image), and obtaining subjects' ratings for both versions. For 77% of the image pairs, we find statistically significant differences in the scores assigned by human subjects, and these differences agree with model predictions.

In essay 2, we use the developed model for predicting visual emotions to study consumer engagement (Retweeting). We collect 1.3 million Tweets from 637 brands from 11 categories, posted on Twitter since 2008. We score sentiment and arousal/motivation-to-act of Tweet text by using existing tools and NLP procedures (e.g., VADER, Hutto & Gilbert, 2014; usage of

exclamation, question marks, uppercase, and "call-to-action" verbs, Barbosa & Feng, 2010). In our regressions of Retweeting on text and image strategies, we control for, among other variables, "brand — week-of-the-year" fixed effects, Tweet-text meaning (approximated by 400-dimensional word vectors), presence of visual (high-level) concepts/objects detected with a deep learning tool (Clarifai.com), embedded text inside an image (if any), and measure of logo color distance (the extent to which a Tweet image resembles the brand logo). We find the effectiveness of high-arousal (activation) text strategy has decreased over time, with its effect switching from positive to negative sometime around 2011. Interestingly, we do not find the same pattern for high-activation imagery. The visual strategy of a high-arousal image with negative sentiment remains dominant over the study period. We suggest potential explanations based on the persuasion knowledge model. These results suggest *marketers should* ***decrease activation through text and instead leverage imagery to activate consumers*** *in their social media communications*. We also detect significant differences in dominant visual strategies by sector (e.g., negative-sentiment images are dominant for charities/non-profits, and positive-high-arousal images are dominant for quick-service restaurants).

Essay 3 examines the effects of the face and gaze of models in the product images on outcomes such as product clicks, orders, and returns. The use of human models in product images (e.g., apparel) is prevalent in both print and digital marketing. Yet, the effects of the model's face and gaze on shopping outcomes have not been examined, particularly in the large-scale empirical setting. We apply deep-learning algorithms for face detection and gaze-following in the context of 57,088 apparel products from 22 categories on one of the largest Chinese e-commerce websites. First, we show people more often browse product images that include the model's face. However, higher prominence of the face leads to less browsing, but more ordering of the product. Second,

we document a *negative impact of the "direct" gaze* (i.e., the model's gaze directed toward the viewer) on both browsing and ordering. The finding is consistent with previous behavioral studies that look at metrics such as dwell time on the product/brand elements of the ad, memorization, and recall. Our study is the first to demonstrate the effects of the "direct" gaze in a large-scale empirical setting. Third, we consider subtypes of the "averted" gaze (i.e., the model's gaze that is not directed toward the viewer), which are "sideways" gaze and "downwards" gaze. Interestingly, *"downwards" gazes result in less browsing and fewer orders than "sideways" (particularly, "to-the-right") gazes*. We suggest a potential explanation for underperformance of the "downwards" gaze stemming from potential associations with shame and embarrassment (Clifford & Palmer, 2018). Taken together, our results suggest actionable guides for product photo designers.

The paper proceeds as follows. We review literature on user- and firm-generated content (UGC and FGC) and the two branches of visual marketing (eye-tracking and computer-vision-based). We then describe details of the three essays, and conclude with an overall discussion and future research directions.

Literature review

L1. User- and Firm-Generated Content. Message Virality.

Since the advent of Web 2.0, research attention has been devoted to the analysis of UGC and FGC. An early investigation by Chevalier and Mayzlin (2006) found that an improvement in a book's user reviews on an e-commerce website increased relative sales. Tirunillai and Tellis (2014) processed high-frequency online chatter in order to infer meaning relevant to consumer satisfaction. Borah and Tellis (2016) found a negative spillover effect of chatter: Negative chatter about a particular product increases negative chatter for its close competition. Mayzlin and Yoganarasimhan (2012) studied competition of web blogs for readership. The authors outlined cases in which blogs benefit from linking to each other despite conventional competition considerations. Netzer et al. (2012) mined brand perceptions from conversations on an automobile forum and came up with perceptual maps/market structures based on users' text UGC.

Some articles focus directly on analyzing FGC. Kumar et al. (2016) found FGC affects consumer spending, cross-buying, and customer profitability. The authors investigated FGC dimensions of valence, receptivity, and customer susceptibility.

Akpinar and Berger (2017) examined ad virality and value (i.e., the ability to generate sales) as a function of emotional versus informative ad appeal. Using a large observational dataset of ad videos, they found a company can accomplish both objectives of content-sharing reach and brand-related outcomes if it uses emotional ads in which the brand is integral. The authors manually coded the variables of informative versus emotional appeals, as well as the degree of brand integrality to the ad.

In another study, Berger and Milkman (2012) investigated the impact of text-content emotionality on content virality. They studied how the emotional content of a news article (*NYT*)

results in the probability of an article going viral. They operationalized emotional content with dimensions of positivity and emotionality (predicted with Linguistic Inquiry and Word Count (LIWC)), as well as the amount of specific emotions such as anxiety, anger, awe, or sadness (manually coded). The authors reported a positive effect of all independent variables on content virality, except for sadness, which had a negative effect. Given that sadness is characterized by lower activation (arousal), the authors suggest high emotional activation (arousal) of the content helps make it viral.

In another empirical study, Lee, Hosanagar, and Nair (2018) used NLP to examine user engagement with brand messages on Facebook. The authors classified brand messages as directly informative (e.g., deals, promotions) and/or persuasive (having brand-personality-related content). They operationalized brand-personality-related content attributes by emotional content, banter, humor, and philanthropic content. The authors found directly informative messages decreased reach in isolation, but increased reach when used in combination with brand-personality-related attributes. Informative content can also increase click-through rates, thus potentially affecting actual sales. The authors hence recommend using some mix of both content types.

Tucker (2014) examined the tradeoff between the reach and persuasiveness of ad videos. Ad attributes included measures of the ad being funny, outrageous, and visually appealing (survey participants assigned the ratings). Although outrageousness can increase the sharing of the ad among users, it decreases persuasiveness. By contrast, humor in an ad can make an ad both viral and persuasive, hence alleviating the tradeoff.

Table L1 presents a summary of the features of previous studies on virality/engagement. In sum, prior research on user sharing of FGC has focused mostly either on text characteristics of the message or on abstract attributes of the message/ad elicited from consumer surveys. Research so

far has not particularly addressed the differential role of the text component and image component of a message. This paper aims to fill this gap and to suggest a machine-learning-based procedure of content annotation that could potentially substitute for the procedure of manual content annotation in future research that deals with measuring content emotionality.

Another characteristic of existing studies on engagement and virality is length (time span of the data). The studies covered a one-year period, on average (two years at most). Our study covers almost 10 years of data on Twitter, which allows us to examine the evolution of the effectiveness of content strategies and uncover valuable perspectives and new insights.

Table L1. Studies Linking Engagement (Virality) to Content Characteristics

Study	**Outcome**	**Main content attributes**	**Approach**	**Image-related variables**	**Time span**	**Main finding / recommendation**
Berger and Milkman (2012)	Probability of *NY Times* article making it to "most emailed" list	Positivity and emotionality + amount of anxiety, anger, awe, sadness	LIWC, Manual text coding	N/A	Aug-Nov 2008	High emotional activation (arousal) of the content helps to make it viral. Presence of low-arousal emotions (sadness) decreases virality.
Lee, Hosanagar, and Nair (2018)	Facebook likes, comments, shares, click-throughs	Informative and/or persuasive messages + brand-personality-related content attributes	Machine learning on text	Photo/video presence in a message	Sep 2011-Jul 2012	Directly informative content should be combined with brand-personality-related content.
Stephen, Sciandra, Inman (2015)	Facebook likes, comments, shares, click-throughs	Arousal, persuasion, information, call-to-action, reference	Manual text coding	Photo/video presence in a message	Mar 2012 – Aug 2013	Adopting a highly persuasive tone hurts engagement with brand messages in social media.
Akpinar and Berger (2017)	Ad shares	Emotional vs. informative ad appeal + brand integratedness	Manual video coding	Ads examined were videos	Jun – Dec 2013	Brands should use emotional ads in which the brand is integral.
Villarroer Ordenes et al. (2018)	Facebook shares, retweets of firm-generated tweets	Speech acts (assertive, expressive, or directive) + figures of speech (alliteration, repetition)	Machine learning on text	Information vs action in the image (manual coding)	Oct 2015 - May 2017	Directive messages are associated with less sharing than both assertive and expressive messages. Images with action decrease sharing.
This study	Retweets of firm-generated tweets	Text and image sentiment/arousal	Machine learning	Sentiment/ Arousal	2008-2017	High-arousal text is associated with less retweeting, and high-arousal imagery can lead to more Tetweeting.

L2. Visual marketing research

L2.1 Eye-tracking research

Extant research in visual marketing focuses on how consumers look at marketing materials (ads, shelves, packaging, etc.) and how visual stimuli affect visual attention and brand/product memory. This research is predominantly based on eye-tracking and behavioral lab experiments. The eye-tracking approach is useful in understanding the cognitive/psychological mechanism behind customers' gaze patterns. The core of the eye-tracking literature focuses on the relationship between ad-design elements and the viewer's visual attention. Wedel and Pieters (2000) examined how eye fixations were distributed across the three elements of a print ad (brand, pictorial, and text). The brand element received the most eye fixations per unit of area. Ads with a larger brand element hence received better recall scores in a follow-up brand-memory test. Pictorials were more effective than text elements in facilitating brand recall. In a larger-scale (1,363 print ads) investigation of the brand, pictorial, and text elements, Pieters and Wedel (2004) examined attention transfer among the elements. Contrary to conventional wisdom, the pictorial element did not facilitate attention transfer to the other two elements: Attention was actually transferred from the brand element. The pictorial element was effective in capturing attention irrespective of the surface size, whereas the text's attention-capture ability was proportional to its surface area.

The way a viewer fixates on different elements of the ad is contingent on the viewer's *goals* (Yarbus 1967). Rayner et al. (2001) found that when an ad was useful for a viewer's goal, the viewer spent more time fixating on this ad compared than on an unrelated ad. Pieters and Wedel (2007) found that an ad-memorization goal increased the length of fixation for all elements of the ad (headline, brand, pictorial, text) compared to free-viewing (no goal). However, under the brand learning goal, fixations shifted significantly from the pictorial to the text element.

Goals also affect how viewers switch between local and global states of visual attention. The global state (Liechty, Pieters, and Wedel 2003) corresponds to a "zoom-out" mode with saccades rapidly moving across the major areas of the stimulus, whereas the local state is more of a "zoom-in" mode characterized by shorter saccades and eyes' detail-oriented focus on one area of the stimulus. Broadly, the global state addresses the question of "where" during a visual scan, and the local state addresses the question of "what" (Wedel & Pieters, 2017). Wedel, Pieters, and Liechty (2008) found that goals affect how often viewers switch between local and global states, and how long they remain in the local (i.e., detail-oriented) state. Like eye fixations, head movements also follow the viewer's goal (Pieters & Wedel, 2018). Under the goal to evaluate the informativeness (vs. attractiveness) of an ad, viewers' head-stimulus distance was smaller (larger), and being closer to the stimulus was predictive of better brand memory.

Another branch of eye-tracking research focuses on the implications of a cluttered visual environment. Pieters, Warlop, and Wedel (2002) found ad originality helps draw consumer attention to the advertised brand in the context of print ads. The best results on brand memory were achieved when an ad exhibited both originality and familiarity. The findings reinforce the benefits of originality and go against the belief that "original ads may win creative awards but lose markets" (p. 777). Pieters, Wedel, and Zhang (2007) offer a model to optimize ad-design elements in order to maximize the impact on consumer attention when multiple distractor ads are present. The authors considered five design elements (brand, pictorial, text, price, promotion) and used the surface area of each element as the optimization lever. They measured competitive clutter using the distinctiveness of the focal ad and heterogeneity of distractor ads. The model suggests an improvement of up to 45% in terms of consumer attention if the surface area is transferred from text and pictorial elements to the other elements.

An important limitation of existing eye-tracking research is that it rarely studies impact on bottom-line business outcomes such as click-throughs, sales, and product returns. A notable exception is Zhang, Wedel, and Pieters (2009). Visual attention (measured as gaze duration) on a feature print ad (1) affects subsequent sales and (2) mediates the impact of specific ad-design elements. Specifically, the color and location of the ad did not significantly affect sales, whereas the display size of the ad did.

Although this evidence of the positive link between visual attention and sales is encouraging, the literature linking elements of visual design to business outcomes remains sparse. Further efforts in establishing a clearer connection between visual attention and business outcomes are encouraged (Wedel and Pieters 2017).

L2.2 Computer-vision approach in marketing

Unlike eye-tracking, the computer-vision approach focuses on the direct link between marketing stimulus design and bottom-line business outcomes such as sales, clicks, product returns, survival rates, and so on. Jalali and Papatla (2016) linked click-through rates of curated user-generated pictures to the color features of these pictures (hue, chroma/saturation, brightness). Higher proportions of green and lower proportions of red and cyan were linked to higher click-through rates. Higher chroma/saturation of red and blue were linked to higher click-through rates. The estimated gains of the dominant visual strategy (i.e., curating content that loads more on features that are linked to higher click-through) were 60% for the clothing category, 73% for mass merchandise, and 110% for discount retail (p. 378).

Liu, Dzyabura, and Mizik (2020) elicited dimensions of brand personality from user-generated brand-related imagery on social media. Low-level visual features in user-generated images portrayed particular traits of brand personality (e.g., glamorous or rugged, fun or dull,

healthy or unhealthy). The authors found that brand perceptions communicated by social media users correlate well with traditional brand-personality surveys such as BAV (e.g., Lovett et al., 2014). Consumer imagery also allows firms to learn unique insights about brand positioning that surveys do not capture.

Zhang et al. (2017) linked attributes of verified photos of Airbnb properties to consumer demand. Verified photos generated over $2500 in additional revenue to an Airbnb host. Importantly, verified photos differed from unverified photos on visual features, even when quality is the same. The computer-vision approach allows to learn the demand-generating visual features and to optimize visual content on these features.

Dzyabura et al. (2018) addressed an important problem of product-return management in online retail. The authors linked return rates to image features of apparel product photos, suggesting consumers might often be ordering non-standard apparel online only to return it in the future. The non-standard apparel would be characterized by, for example, the presence of pink, striped/checkered texture versus standard blue, brown, or black. Including visual features in a predictive model for the product-return rate improved accuracy by 10% over the baseline, suggesting a considerable value of visual information. Incorporating image data into an optimal policy of product-launching resulted in an 8.5% increase in profitability.

Zhang and Luo (2018) linked customer review photos from Yelp to restaurant survival rates. Photos were more accurate in predicting survival than text reviews, and the prediction horizon of photos was longer than text reviews (three years and one year, respectively). Hence, customer photo reviews serve as valuable information for capital-allocation decisions in the restaurant business.

Lu, Xiao, and Ding (2016) proposed an automatic apparel recommender system for in-store purchases. The system applied computer vision to video frames of live apparel customer try-ons. The algorithms allow the extraction of, for example, customer interactions with the piece of clothing (hand positions over the regions of a garment) and customer satisfaction from facial expressions. After observing the satisfaction level, recommendations are made based on similar customers' past shopping choices. The approach was an early and innovative way to apply computer vision to video data for marketing applications.

Ho (2017) extended traditional hedonic models in economics (used to model property values as a function of observable factors) with visual features of property photographs on hosting platforms such as Zillow.com. Ignoring these features in hedonic regressions can result in a major omitted-variable bias, for example, more than a 50% underestimation of the effect of air pollution on property prices.

Although promising, the computer-vision applications in marketing are still nascent. This dissertation aims to contribute to the computer-vision branch of visual marketing by (1) examining the relative impact of text and image emotions in firm-generated social media on user engagement and (2) examining the effects of the model's face and gaze in a product image on product browses, orders, and returns.

Essay 1. Visualized Emotions: A Model for Extracting Emotional Loading of Imagery

1.1. Theoretical Frameworks for Emotions

The tradition of characterizing emotions along certain attributes/dimensions traces back to Duffy (1934). Prominent dimensional models of emotion include the circumplex model (Russel, 1980), the Positive Activation – Negative Activation (PANA) model (Watson & Tellegen, 1985), and the vector model (Bradley et al., 1992). All three models employ dimensions of (1) sentiment/valence (pleasant-unpleasant) and (2) arousal/activation (low-high), Figure 1.1. The dimension of arousal/activation not only reflects the intensity of a particular sentiment, but also helps distinguish between different emotions that have comparable valence. For example, both "angry" and "sad" have negative sentiment. Yet, the former is a high-arousal emotion, whereas the latter is low arousal.

The *circumplex* model posits that emotional stimuli are located inside a circular region centered at the origin. The *vector* model suggests the impossibility of "intensely neutral" stimuli, that is, ones with close-to-zero sentiment and high arousal. It suggests that high levels of activation could be experienced only at the extreme levels of sentiment. Rubin and Talarico (2009) provide evidence that the vector model outperforms the circumflex model in the goodness of fit of subjects' ratings of emotional stimuli along sentiment and arousal dimensions. Also, the vector model has been specifically employed in the analysis of visual stimuli (Bradley et al., 1992), which is the focus of this paper. Given these two observations, we used the vector model of emotion as our organizing framework.

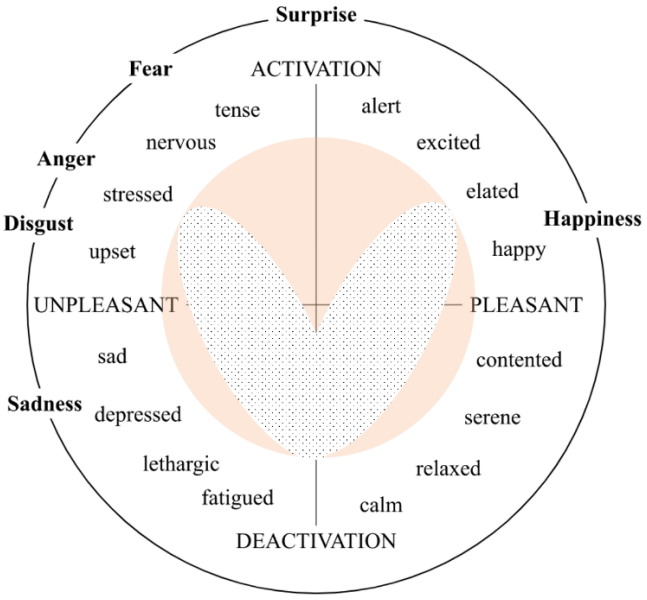


Figure 1.1. The Pleasantness-Activation Structure of Emotion

Note. adapted from the Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/emotion/. The red region denotes the possible location of emotional stimuli according to the circumplex model (Russel 1980); the dotted region denotes the possible location of emotional stimuli according to the vector model (Bradley et al., 1992). The two regions are added by the author following Rubin and Talarico (2009).

Our goal was to develop a model that automatically puts an image on the dimensions of sentiment and arousal (activation). Such a model requires a dataset of images that human subjects pre-labeled on sentiment and arousal dimensions. We collected such a dataset using the behavioral lab, which we explain next.

1.2. Training set of images

We utilized the behavioral lab to annotate the training set of images on sentiment and arousal dimensions. Undergraduate UW students participated in the survey in exchange for a course credit. The study was conducted during four quarters (Fall 17, Spring 18, Fall 18, Winter 19). A total of 1292 subjects participated across all quarters. A total of 13,386 original brand-generated images from twitter were annotated. The images were randomly sampled from a pool of 1.3M images from 637 brands on Twitter. We obtained the brand list from the Young and Rubicam Brand Asset

Valuator (BAV, Lovett et al., 2014). The 637 brands represent 11 unique categories (both profit and nonprofit); see Table 1.1. The time range of Tweet images was from 2011 to 2017, with a median image coming from 2015 (see Table 1.2).

Table 1.1. Categories of the Training Set of Images

Category	N img	Freq, %
Apparel & Accessories	3,717	27.77
QSR (quick-service restaurants)	1,509	11.27
Auto	1,366	10.21
Nonprofit/charities	1,243	9.29
Health and Beauty	1,241	9.27
Travel & Entertainment	1,229	9.18
Food	1,055	7.88
Beverages - Non Alcoholic	616	4.6
Beverages - Alcoholic	561	4.19
Household Products	488	3.65
Other	360	2.7

Table 1.2. Year Range of the Training Set of Images

Year	N img	Freq, %
2011	41	0.31
2012	461	3.44
2013	1,254	9.37
2014	2,864	21.4
2015	3,517	26.27
2016	3,250	24.28
2017	1,999	14.93

The sequence of images shown to the subjects was randomized. When shown an image, the subjects were asked two binary-response questions, the first one measuring sentiment and the second one measuring arousal:

1. “Does this image make you feel optimistic (positive emotions) or pessimistic (negative emotions)?”

2. “Does this image make you feel stimulated or relaxed?”

Figure 1.2 shows an example of a stimulus and the questions

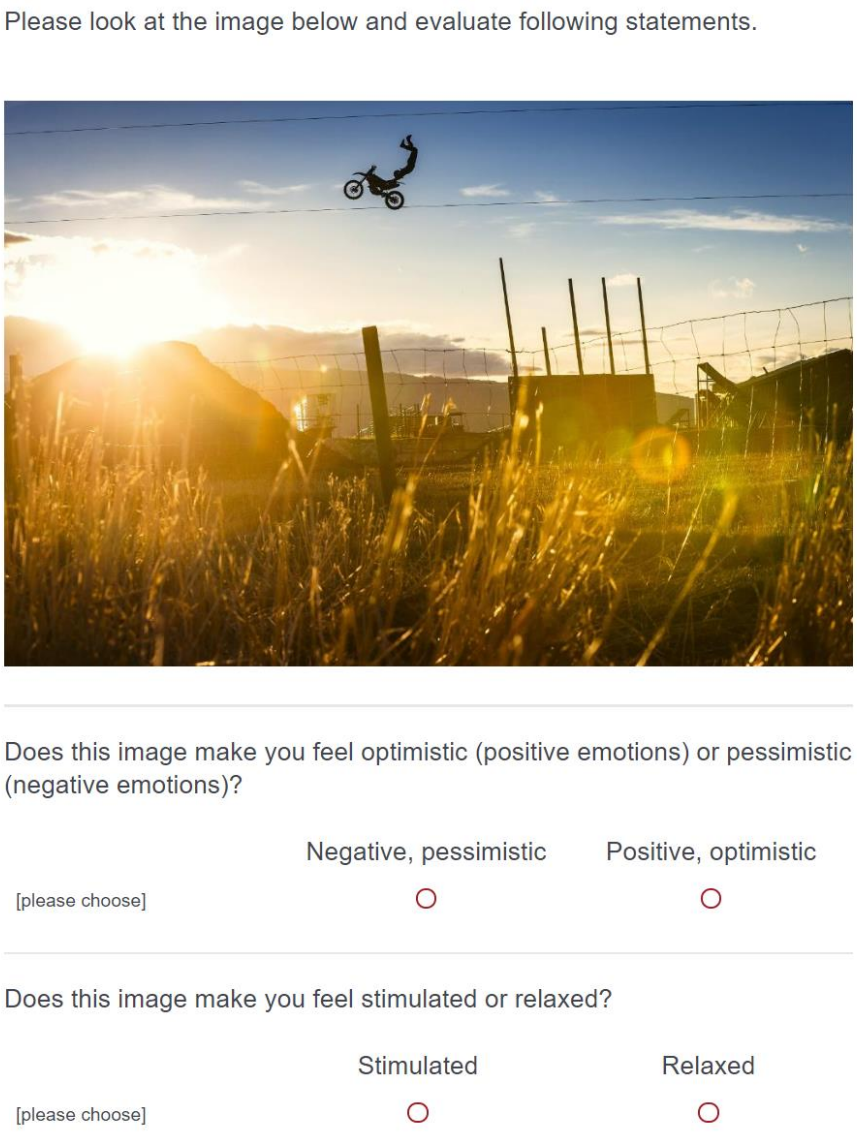


Figure 1.2. Example Stimuli and Questions for the Lab Annotation Task

In terms of the number of subject evaluations per image, the median image received 5 evaluations, and an average image received 7.6 evaluations. Individual binary responses were then aggregated for each image as a summation of "-1" and "+1" evaluations. For example, if an image received 5 evaluations with 3 being "positive" and two being "negative," the aggregate sentiment score for the image is 3-2=+1. Under this design, images that (randomly) received more

evaluations would score higher on sentiment even if the relative vote is the same. For example, if an image obtained 15 evaluations, with 9 being "positive" and 6 being "negative" (i.e., the same relative vote as in the previous example), the aggregate sentiment score for the image is 9-6=+3. This approach ensures that an image can have extreme levels of sentiment only if a substantial number of subjects saw this image (i.e., the subjects' consensus is more reliable).

The scatterplot of ground-truth visual sentiment and arousal is in Figure 1.3. Because the zero line on both of the axes is informative (i.e., if an image locates to the right from the "Sentiment=0" line, subjects' consensus vote on this image is predominantly positive), we can gauge the relative proportions of images by the emotional quadrants. Twenty-eight percent of images belong to "positive, high-arousal"; 53%, to "positive, low-arousal"; 12%, to "negative, high-arousal"; and 7%, to "negative, low-arousal." Hence, the largest share belongs to the positive, low-arousal segment. However, we have a substantial presence of negative sentiment as well, particularly in the negative, high-arousal quadrant. Figure 1.4 presents a scatterplot of average ground-truth visual sentiment and arousal by category. We detect interesting category differences, such as more arousing and less positive images in the Nonprofit sector, and very positive and low-arousing images in the Travel & Entertainment sector.

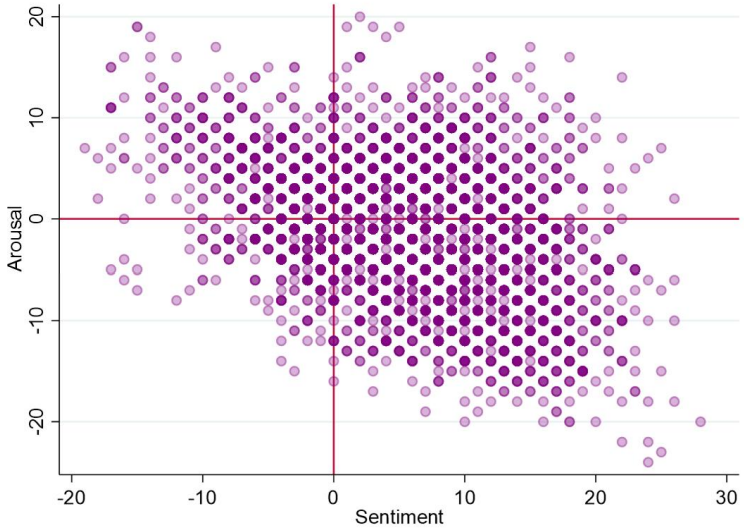


Figure 1.3. Scatterplot of Ground-Truth Visual Sentiment and Arousal (N=13,386)

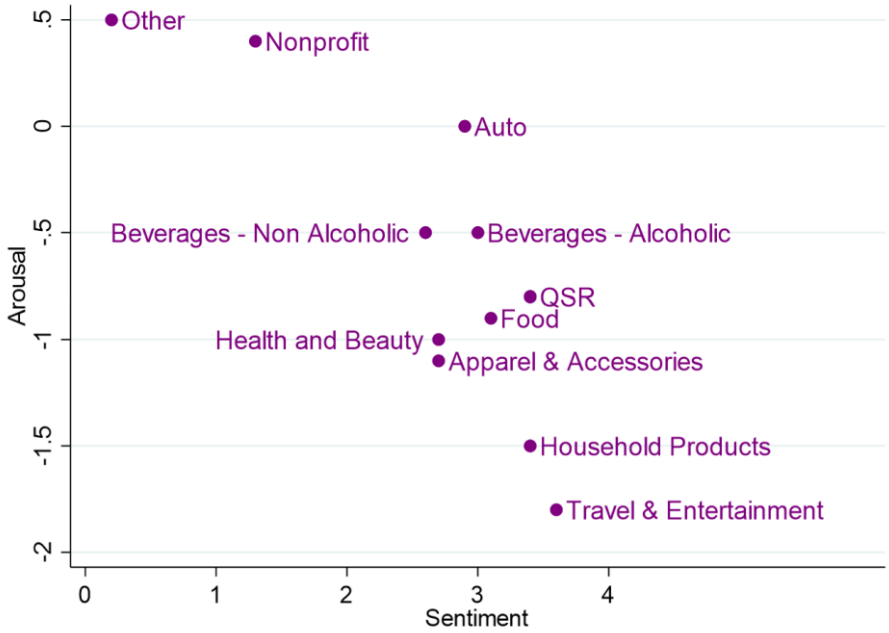


Figure 1.4. Average Sentiment-Arousal Scatterplot by Category

Figure 1.5 displays an example of annotated images from the training set, by the four emotional quadrants (sentiment [negative, positive] × arousal [low, high]). The images appear to differ by quadrant based on color properties (e.g., brightness, color variety, presence of individual hues such as red, green, pink), by objects present (e.g., beach vs. girl vs. fire vs. trash), and by

other characteristics. We next systematically examined visual-emotion modalities in an attempt to identify the drivers of visual emotions.

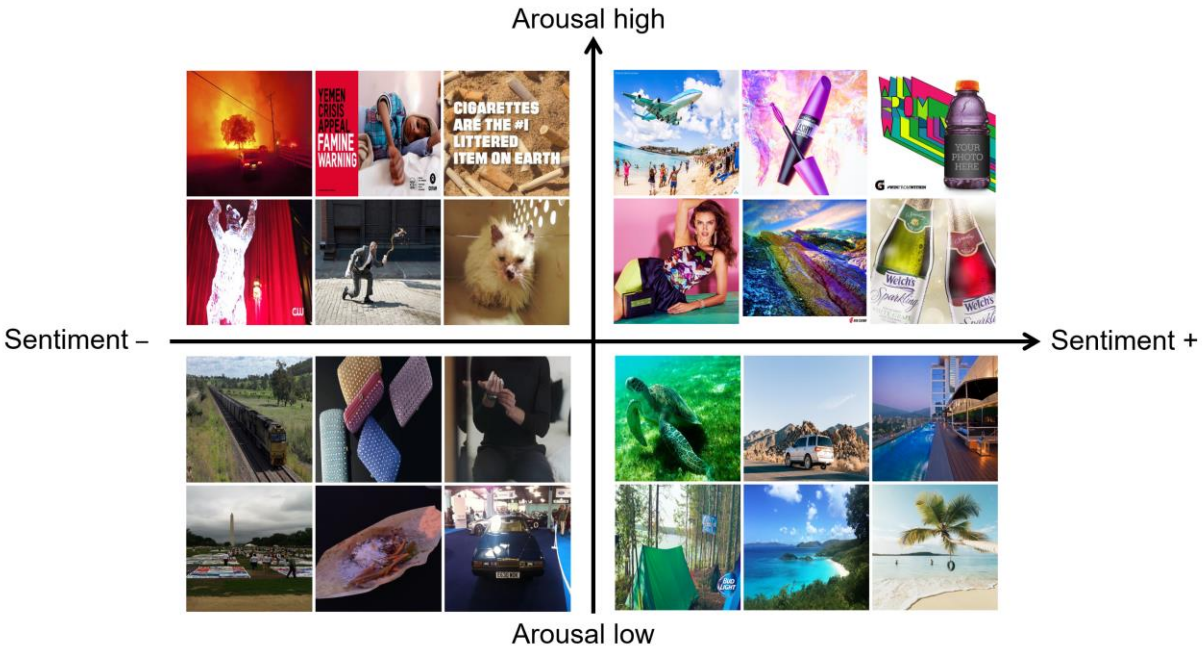


Figure 1.5. Example of Annotated Images from the Training Set, by Emotional Quadrant

1.3. Visual-emotion modalities

Possessing a ground-truth annotated set of training images, we next developed a model that links attributes of the images to their sentiment/arousal labels. We propose a multi-modal sentiment/arousal prediction model (see Figure 1.6). The four modalities include (1) visual features – fundamental elements of design (color, texture, shape, lines, curves, corners, edges, orientation), (2) high-level visual objects/concepts (e.g., adventure, action, leisure, danger, etc.), (3) human facial expressions, and (4) text embedded in the image. Our eventual goal was to extract numerical variables characterizing each of the four modalities, and use these variables to predict the numerical sentiment/arousal score of an image. We next describe each of the modalities in detail.

Figure 1.6. Visual Emotion Modalities

1.3.1. Modality 1: Visual features – fundamental elements of design

To be able to link a digital image to its numerical Sentiment/Arousal level, we need to represent this image as a set of quantifiable characteristics (elements of design). Although different conceptualizations exist, most design textbooks (e.g., Hashimoto & Clayton, 2009; Dondis, 1974; Arnheim, 1965) consistently identify color, shape, and texture as the fundamental elements of visual design. Like an artist who organizes elements within a work of art, a designer uses design elements to achieve his design goals. To a visual designer, the elements of design are the building blocks used to create a piece of design. We used elements of design relevant to the analysis of photography aesthetics (Khosla et al., 2014; Datta et al., 2006) and image likability (Machajdik & Hanbury, 2010; Buter et al., 2011).

Color is perhaps the most complex of the elements of design, and many theories exist about the meanings and emotional associations of color. For example, lighter colors suggest a brighter, happier mood, whereas darker values feel somber and serious. Oranges, reds, and yellows are viewed as warm, and blues as cold. Red is considered exciting and energetic; yellow, optimistic and upbeat; green, refreshing; and blue, cool. These associations are believed to come from the individual's observations of the natural world and may differ across cultures.

Munsell's (1912) color system, developed over a century ago to represent color in a rational/quantifiable way, remains the most popular color-representation system in art and design. Munsell developed and continued to improve his color system based on rigorous measurements of human perceptions and visual responses to color. Color has three main properties distinguishable by the human eye: *hue, saturation,* and *brightness* (HSB). These three properties constitute the three independent dimensions in the Munsell system, and any color can be specified with three numbers for these dimensions. Hue refers to the pure state of a color. We used hues as the names of colors, such as red, blue, or yellow. Saturation (or intensity, strength, purity, chroma) refers to the vividness of the color. High saturation means the color is close to its pure hue, whereas a hue with 0% saturation appears gray. Brightness (also known as lightness, tone, value) reflects how light or dark it is. Brightness measures the relative degree of black or white mixed with a given hue. It captures the lightness or darkness of colors and runs from white to black.

An image is a collection of pixels coded by a certain number. In any digital image, each pixel is a 3-tuple, each tuple representing the pixel's score on a color property. HSB contains hue (color index from the pallet of possible colors, range 0-180), saturation (whether the color is subtle or highly saturated, 0 and 1, respectively), and brightness (dark vs. light, 0 and 1, respectively). Color values are summarized through color histograms. Consider hue. We split the possible range of hues (0-180) into 20 bins and counted how many pixels of an image fall into each bin. Figure 1.7 presents a sample image and normalized color histograms of RGB (red, green, blue) and HSB. Among the HSB histograms, hue (red line) is concentrated in bins 9-12 (orange bins), and saturation is bi-modal with a bigger cluster massed around zero (corresponding to grey pixels that have low saturation) and a smaller cluster massed around 16 (corresponding to orange pattern). Brightness is massed around bins 17-18, reflecting that image is generally light. RGB histogram

appears less informative—all RGB channels are concentrated around 1 (which corresponds to a large overall proportion of white in the image) and appear to have much higher correlation than HSB histograms.

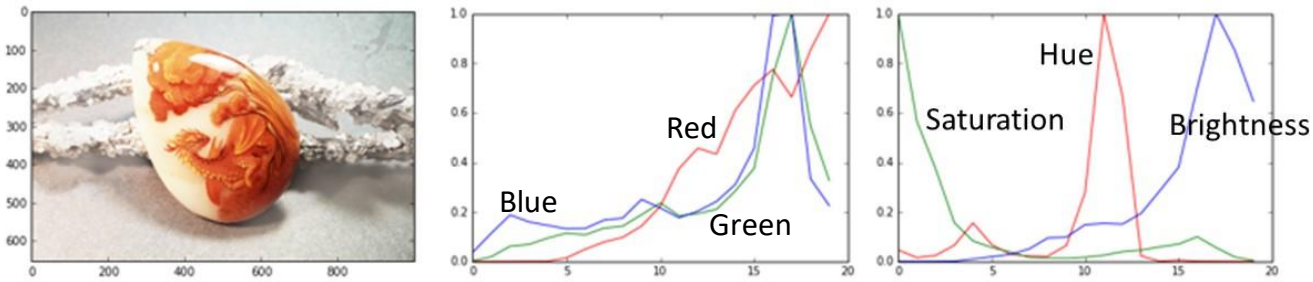


Figure 1.7. Sample Image and Its RGB and HSV Histograms.

Note. The left panel presents a sample image. The center image represents the RGB color histogram, normalized. The right panel represents the HSB color histogram, normalized (red=hue, green=saturation, blue=brightness).

Figure 1.8 displays more examples of images with different profiles of HSB histograms. The Fox News logo has only one color that corresponds to single spikes in hue, saturation, and brightness. The CNBC logo has six color clusters, and each of them corresponds to a local spike in hue, saturation, and brightness. Finally, egg art has more complex hues, gradients, and shades, which corresponds to more continuous HSB histograms.

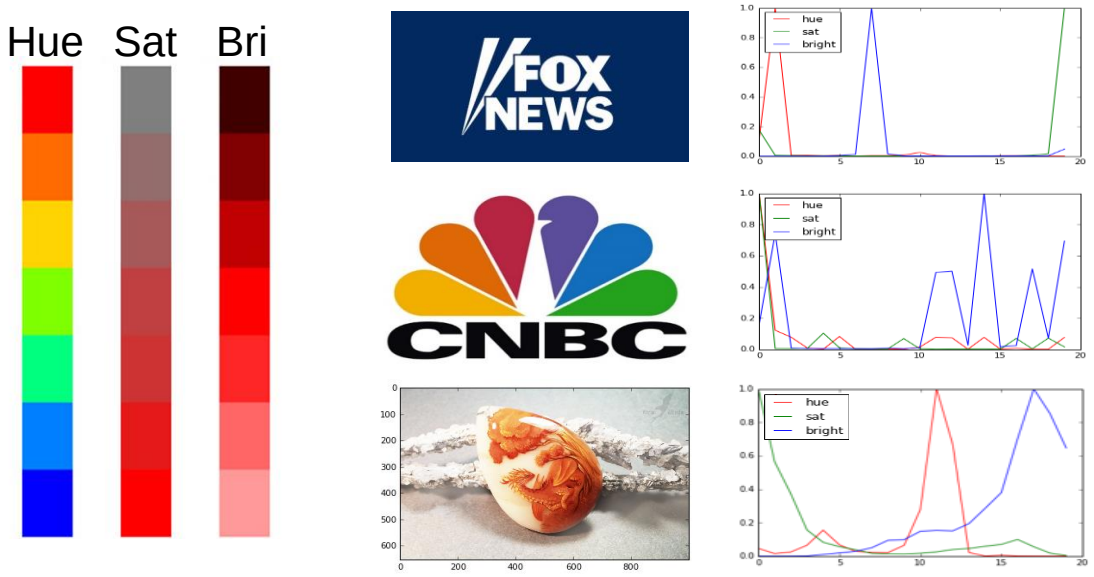


Figure 1.8. Examples of Images with Different Profiles of HSB Histograms

Note. HSB stands for hue, saturation, and brightness.

Additional variables that can be created based on HSB/RGB pixel code include the mean, standard deviation of each channel, and coefficient of concentration defined as $HHI = \sum_i^{20} \alpha_i^2$, where α_i is the proportion of the i-th bin of a given channel across all bins of this channel. The sum of alphas is always 1. Higher HHI means a higher concentration of pixels in a single bin. For hues, it would imply the presence of a dominant color (or color tonality). A lower HHI of hue means a more diverse image in terms of the color palette. Color variety is hence the inverse of the HHI of hues. Figure 1.9 shows example images that exhibit the presence of a dominant color or, on the contrary, color variety.

Figure 1.9. Sample Images Sorted by Visual Feature

Note. Groups 1, 2, 3, and 6 of images are sorted based on the prevalent hue. Images from these groups score highly on the color-concentration index (HHI on respective hue). Group 4 scores low on tonality

because it is characterized by high color variety. Group 5 (black and white) scores 0 on color saturation and color variety.

Figure 1.10 presents examples of images with varying brightness and saturation levels. Brightness measures the overall amount of light in the pixels. Saturation measures hue intensity in the pixels. Black-and-white images will have saturation equal to zero.

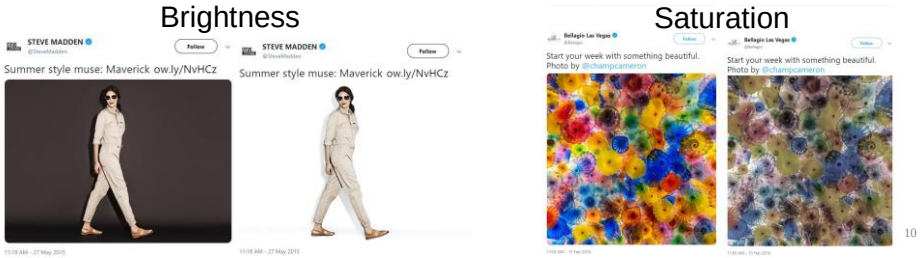


Figure 1.10. Examples of Images with Varying Brightness and Saturation Levels

Texture captures the surface quality of an object, which would be sensed through touch. Texture depicted in an image causes the viewer to imagine the sensation experienced when touching the surface. We used the Canny algorithm to assess the roughness/smoothness of texture by identifying number of edges in the image. Consider Figure 1.11: A lower number of detected edges corresponds to a smoother texture.

Shape has several building blocks: lines, corners, and orientation. Lines are the most basic component of shape. Lines can be thin and delicate or thick and strong, curved and organic or sharp and mechanical. An important characteristic of a line is its direction. Vertical lines suggest power and stability, like a soldier standing at attention. Horizontal lines likewise create a sense of stability, but unlike vertical lines, they also provide a sense of calmness. For example, landscapes and seascapes stretching out horizontally are relaxing and soothing. Shape is a critical cue to identifying and recognizing objects in an image, but it also conveys cognitive, symbolic, and perceptual meanings beyond the surface appearance. We quantified corners by extracting edge-to-pixel, corner-to-pixel, and corner-to-edge ratios. Corners in the image are identified with the Harris

algorithm. An example of corner detection is in Figure 1.12. For vertical/horizontal orientation of lines, calculate the variance of Sobel / Laplacian gradients (Figure 1.13). To calculate 45-degree / 135-degree orientation, we rotated the image before extracting the Sobel / Laplacian gradients (Figure 1.14).

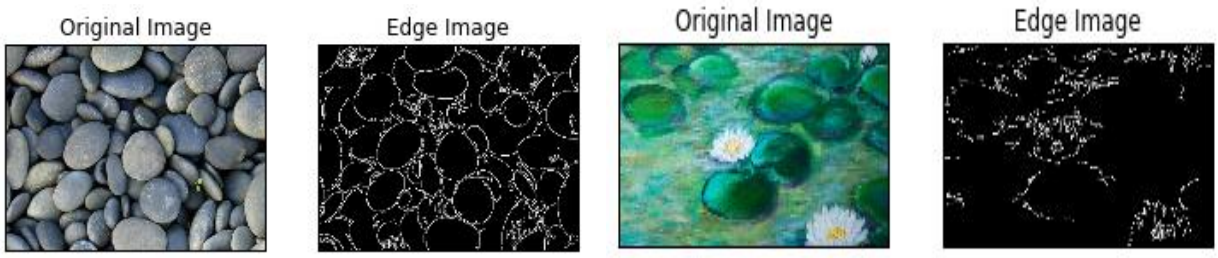


Figure 1.11. Edge-Detection Algorithm Example (Based on the Canny Algorithm)

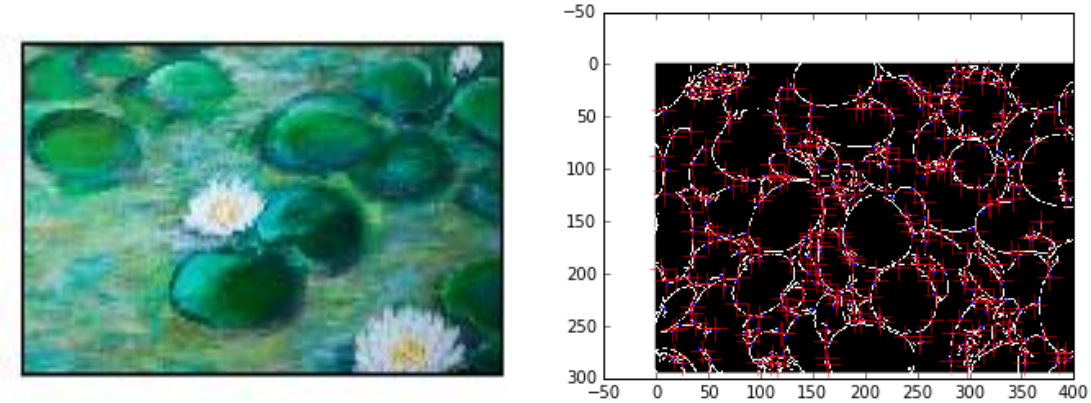


Figure 1.12. Corner-Detection Examples (Harris Algorithm)

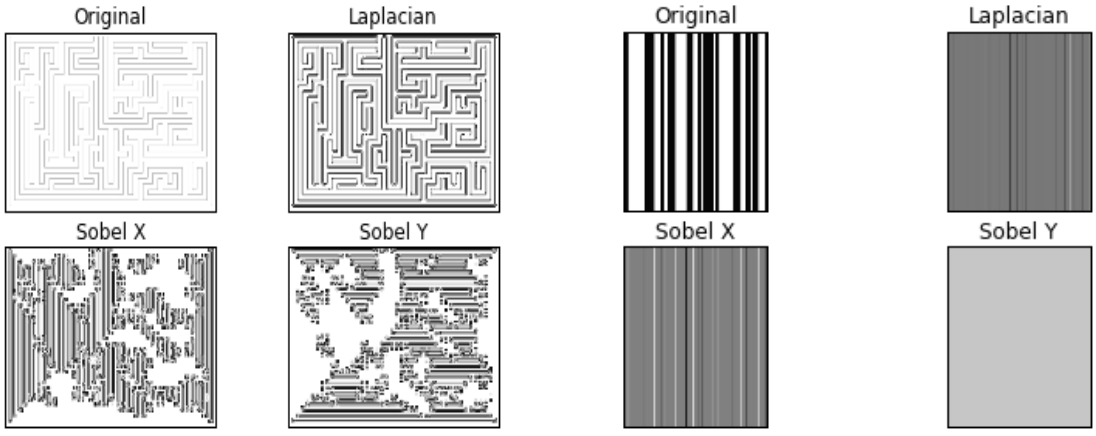


Figure 1.13. Line Orientation.

Note. The four images on the left represent no distinct vertical or horizontal orientation. The four images on the right represent distinct vertical orientation (measure of vertical orientation: 2544; measure of horizontal orientation: 0.371).

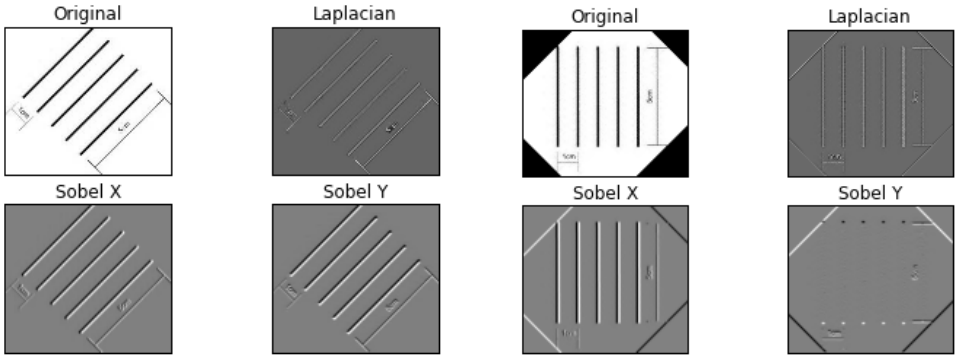


Figure 1.14. Line Orientation.

Note. The four images on the left represent 45-degree image orientation. The four images on the right represent rotated versions with the application of vertical/horizontal filters. Resulting metrics: vertical orientation, 1,966; horizontal orientation, 195; 45-degree orientation, 2,815; 135-degree orientation, 1,126. Hence, the algorithm finds a 45-degree image orientation most likely.

Having described the visual features that we extracted from images, we proceed to describe other modalities of visual emotion.

1.3.2. Modality 2: High-level objects/concepts

Images may portray high-level objects/concepts that have semantic meaning and, potentially, emotional associations. Consider, for instance, Figure 1.15. It portrays a mountain bike sport that communicates concepts of adventure, action, recreation, motion, leisure, and danger, which might put an image into the positive, high-arousal emotional quadrant. Hence, the high-level objects/concepts represent an independent emotional modality.

Concept	Prob	Concept	Prob
bike	98%	motion	92%
adventure	98%	sport	92%
trail	96%	soil	91%
biker	96%	wheel	90%
hurry	96%	man	90%
helmet	95%	road	89%
travel	95%	people	88%
mountain	94%	seated	88%
action	93%	leisure	88%
recreation	93%	danger	87%

Figure 1.15. High-level Objects/Concepts in an Image.

Note. The concepts extracted using a deep-learning tool, Clarifai. Probability denotes the certainty associated with a particular object/concept.

We used an existing deep-learning tool, Clarifai.com, to extract high-level objects/concepts from the images. The tool extracts 20 concepts from each image and predicts the probability (certainty) for each concept. Overall, we detected 3,516 unique concepts in the entire training set of 13,386 images. The distribution of concept probabilities is depicted in Figure 1.16. If we only look at concepts that have higher-than-median probability (0.917), we are left with 2,919 unique concepts. We identified 1,528 unique concepts in the top-decile probability (0.983). Table 1.3 presents most frequent concepts whose probability is at least 97%, by category. We observe that many objects are clustered within a category. In fact, for some objects, brand fixed effects explain up to 60% of the variation in the object occurrence (e.g., “car” or “vehicle”).

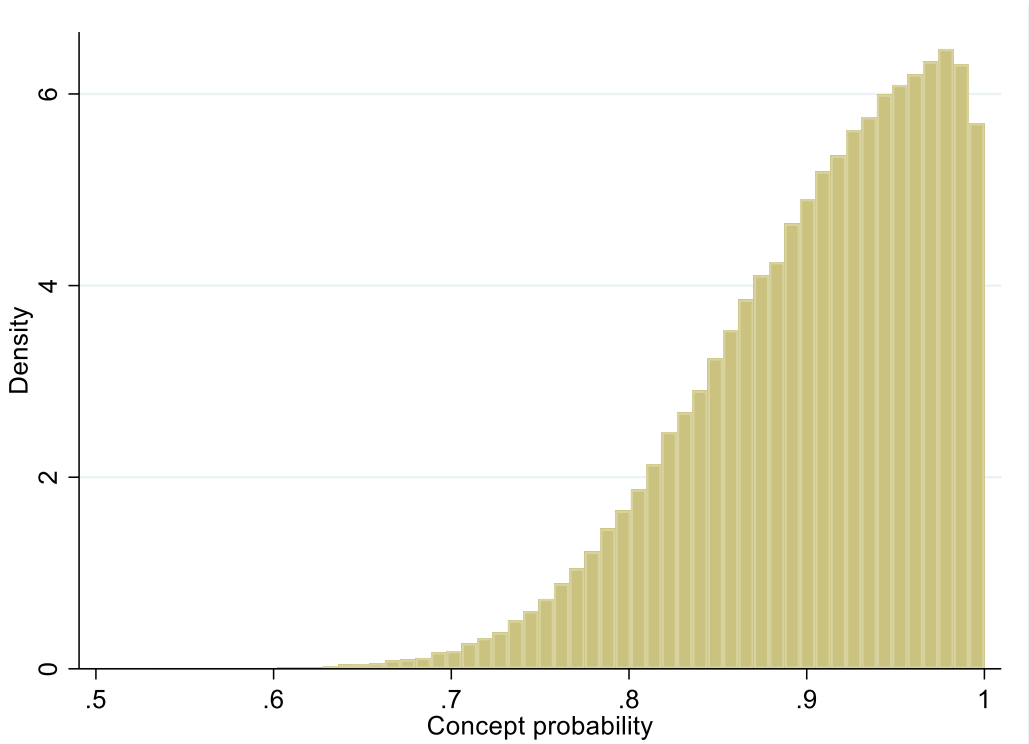


Figure 1.16. Distribution of Concept Probabilities Based on 3,516 Unique Concepts in 13,386 Images

Table 1.3. Top-10 Objects/Concepts by Category

Category	Top 10 objects/concepts
Apparel & Accessories	people, woman, fashion, adult, wear, no person, one, man, portrait, business
QSR (quick-serv rest)	no person, food, delicious, dinner, meal, lunch, refreshment, plate, cooking, nutrition
Auto	vehicle, car, transportation system, drive, fast, wheel, no person, road, automotive, hurry
Nonprofit/charities	people, adult, man, woman, outdoors, business, portrait, no person, child, indoors
Health and Beauty	no person, people, business, woman, adult, desktop, indoors, fashion, portrait, paper
Travel & Entertainment	no person, travel, outdoors, people, sky, water, indoors, business, architecture, adult
Food	no person, food, delicious, desktop, refreshment, business, nutrition, meal, sweet, traditional
Beverages - Non alcoh	no person, people, food, desktop, outdoors, business, adult, man, health, indoors
Beverages - Alcoholic	no person, drink, people, glass, indoors, food, bar, adult, bottle, man
Household Products	no person, cute, animal, portrait, mammal, pet, little, sit, looking, one

Whether objects/concepts in our dataset indeed carry emotional associations is of interest. As an illustration, we assigned scores of sentiment and arousal to our visual concepts using an existing dictionary of emotional keywords, “Affective Norms for English Words” (ANEW, Bradley & Lang, 1999). The ANEW database is one of the standard lexicons used in (text) sentiment

classification (e.g., Corona et al., 2015). ANEW represents a survey in which respondents annotated 1,034 English words along dimensions of sentiment, arousal, and dominance. For instance, “miracle” received scores of 8.6 on sentiment and 7.7 on arousal. We use an updated dictionary (Warriner et al., 2013) that extends the set of annotated words from the original 1,034 to 13,915.

Figure 1.17 shows a sentiment-arousal scatterplot of our objects/concepts superimposed over the original ANEW word cloud. We observe a significant overlap, particularly in the region of positive sentiment. Hence, objects/concepts have considerable variability in the sentiment-arousal levels they carry. Table 1.4 presents the top-10 objects/concepts associated with the highest/lowest sentiment/arousal, as well as the top-10 most neutral concepts.

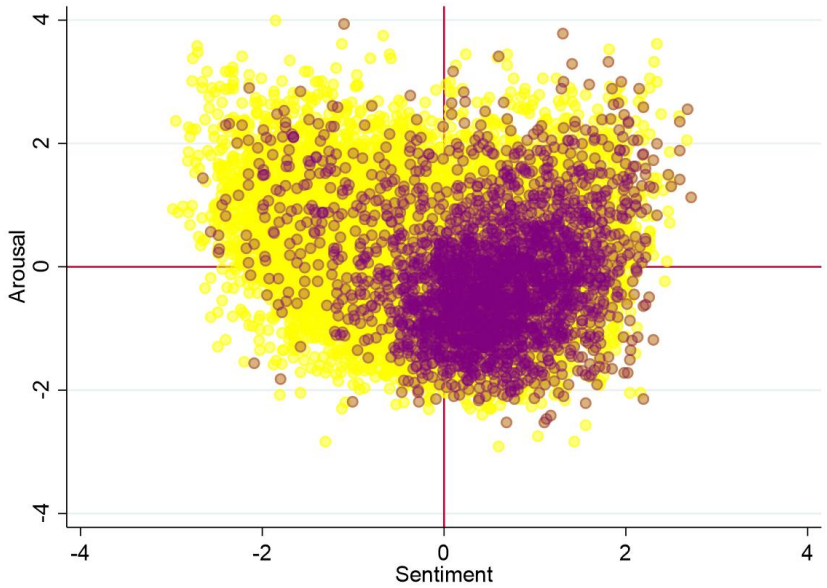


Figure 1.17. Overlap of the Two Word Clouds: Objects/Concepts and the ANEW Dictionary.

Note. Yellow denotes the ANEW 14k word cloud (Warriner et al., 2013). Purple denotes the objects/concepts present in our training data.

Table 1.4. Most Emotional and Most Neutral Objects/Concepts in the Training Set

Dimension	Top 10 objects/concepts
Most positive (+)	vacation, happiness, Christmas, fun, enjoyment, free, joy, comedy, accomplishment, cheerful
Most negative (–)	disease, stress, vandalism, overworked, jail, illness, infection, pollution, pain, crisis
Most arousing (↑)	gun, sex, erotic, thrill, seduction, intensity, exotic, championship, money, dangerous
Least arousing (↓)	scene, quiet, asleep, tea, blanket, cardboard, empty, broth, rest, mailbox
Most neutral (0)	cavalry, cluster, extension, horn, industry, infrastructure, microphone, platform, slick, strap

Note. "Most Neutral" is calculated as a minimum quadratic distance from the origin in a sentiment-arousal plane.

For our subsequent analysis, we retained the objects that (1) have at least 97% probability (/certainty) and (2) appear in the entire dataset at least 10 instances. This approach left us with 583 unique concepts. We created a dummy variable for each of them.

1.3.3. Modality 3: Emotional expression of the faces

Another emotional modality is facial expressions. Facial expressions may carry emotional associations. We used the deep-learning tool "Paralleldots.com" to detect facial emotions. The tool is able to extract seven distinct emotions (happy, angry, neutral, disgust, surprise, sad, and fear) and output a certainty measure for each of them. Figure 1.18 presents prediction results for two images from our training set.

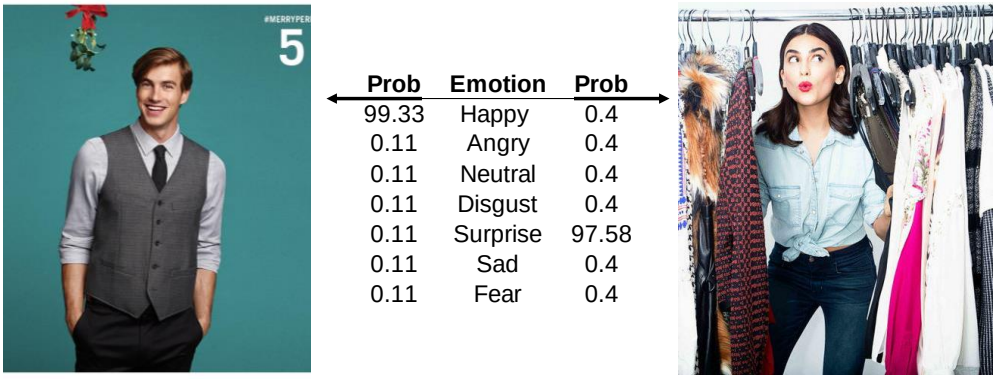


Figure 1.18. Facial Emotion Extraction Using the Deep-Learning Tool Paralleldots.com

Importantly, not every image in our sample hds a person/face present. Per Table 1.3, the "no person" concept was common in many categories. We detected at least one face in only 18.8% of the images. Certain images had multiple faces, and the facial-emotion-extraction algorithm is robust to multiple faces in an image. The frequency of facial emotions conditional on the face being present in an image is in Table 1.5.

Table 1.5. Frequency of Facial Emotions (Conditional on Face Present in an Image)

Variables	N	Mean	SD	Min	Max
Angry	2,518	0.0680	0.0963	0	0.818
Disgust	2,518	0.0362	0.0334	0	0.143
Fear	2,518	0.0558	0.0742	0	0.770
Happy	2,518	0.219	0.318	0	1.000
Neutral	2,518	0.200	0.249	0	0.992
Sad	2,518	0.0764	0.0985	0	0.655
Surprise	2,518	0.0488	0.0701	0	0.976
# faces	2,518	1.675	1.498	1	29

Note. # faces denotes number of faces present in an image (for cases of group photos). N=2,518 images out of 13,386 have at least one face.

We retained all seven facial emotions along with the "# faces" variable for subsequent analysis. If a face is not present in an image, all eight variables are equal to 0.

1.3.4. Modality 4: Embedded text

The fourth emotional modality is text embedded in the image (if any). We used the optical character recognition-based text extractor "PyTesseract" to automatically extract embedded text strings from an image. We detected embedded text in 25% of the images. Next, we analyzed the extracted text strings in terms of text sentiment. We used the "Valence Aware Dictionary and sEntiment Reasoning" (VADER) optimized specifically for Twitter posts (Hutto & Gilbert, 2014) as a sentiment classifier. The algorithm uses annotated vocabulary to assess Tweet-text valence on a scale from -1 to 1. It incorporates information from emoticons and takes into account the intensity of a particular sentiment (e.g., ":)))" will have a more positive sentiment than ":)"). Figure 1.19

shows two examples of the opposite text sentiment. Figure 1.20 graphs the histogram of embedded text sentiment conditional on a text string being present in an image.

Text Sentiment: **-0.9246** (negative) Text Sentiment: **0.9265** (positive)

Figure 1.19. Examples of Sentiment Annotation for an Embedded Text

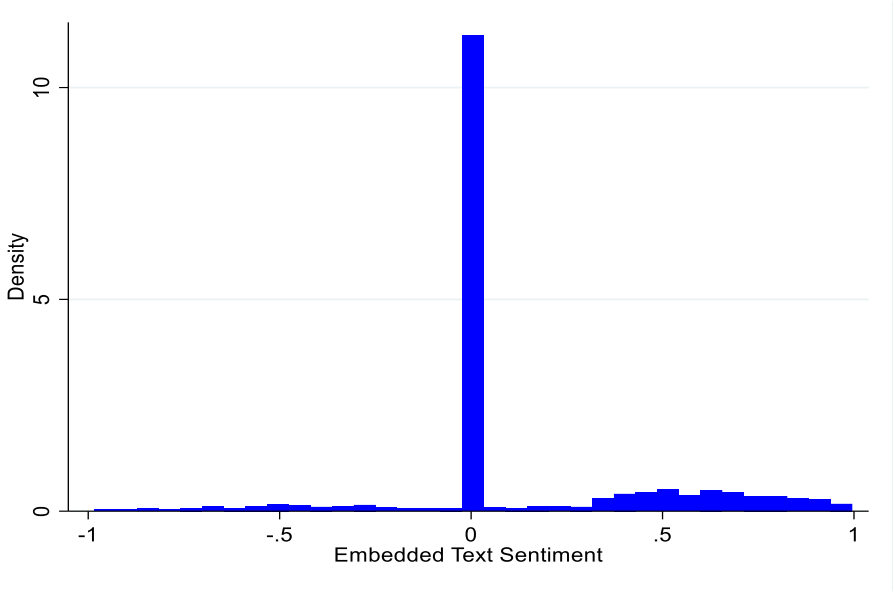


Figure 1.20. Histogram of the Embedded Text Sentiment

Along with the embedded text sentiment, we extracted other characteristics of the text strings: text string length, punctuation, capitalization, presence of call-to-action words (ask, come, try, share, buy, save, grab, invite, enjoy, etc.), presence of promotional keywords ("% off," "discount," "sale"), and number of hashtags. Note the Twitter character limit (140 characters before Nov 7,

2017, and 280 after) does not apply to pixel text, allowing the length of the embedded text to exceed 2,000 characters.

Table 1.6. Embedded Text Variables

Variables	N	Mean	SD	Min	Max
Embedded Text Sentiment	3,452	0.124	0.340	-0.986	0.996
Text string length	3,452	77.98	141.8	1	2,115
"!" mark	3,452	0.0904	0.328	0	4
"?" mark	3,452	0.0513	0.263	0	5
Uppercase characters	3,452	20.06	38.47	0	1,452
"Call-to-action" words	3,452	0.196	0.617	0	12
"% off" keyword	3,452	0.00666	0.0814	0	1
"Discount" keyword	3,452	0.00492	0.149	0	6
"Sale" keyword	3,452	0.0154	0.196	0	8
Hashtags	3,452	0.0857	0.352	0	7

Note. N=3,452 images out of 13,386 have embedded text.

We retained all variables from Table 1.6 for the subsequent analysis. If no embedded text was detected in an image, all variables are equal to 0.

1.4. Visual-emotion prediction model

We used the variables associated with the four modalities in order to predict sentiment and arousal levels for the images from our ground-truth set. After estimating the model with the best accuracy, we used it to predict visual emotions out of sample (i.e., for images without ground-truth sentiment and arousal annotations). Equation (1.1) describes our predictive model for an image i:

$$Y_i = c + Visual\_Features_i + Concepts_i + Face\_var_i + Embedded\_text\_var_i + e_i \quad (1.1)$$

Here, for image i,

- Y_i is the dependent variable. We estimate separate models for sentiment and arousal.
- c denotes the intercept.
- $Visual\_Features_i$ include 114 variables measuring visual features (HSB, RGB channel bins, concentration of hue, saturation, brightness, five principal components for HSB and

RGB channels, measures of corners and edges, measures of horizontal/vertical, and 45-/135-degree orientation).

- $Concepts_i$ include 583 dummies for visual objects/concepts. Each dummy is equal to 1 in at least 10 instances in the data, and the concepts have at least 97% certainty of being present in the image.
- $Face\_var_i$ include face-presence indicator, number of faces present, and seven variables measuring the probability of occurrence of the possible facial expressions (happy, angry, neutral, disgust, surprise, sad, and fear). If the image does not contain a face, all variables are equal to 0.
- $Embedded\_text\_var_i$ include indicators of embedded text presence, string-text length, text sentiment, usage of punctuation ("!","?"), call-to-action words, promotional keywords ("sale," "% off," "discount"), number of uppercase characters, and number of hashtags (total of 10 variables). If the image does not contain embedded text, all variables are equal to 0.
- e_i denotes an idiosyncratic error term.

We followed a common approach to estimating a predictive model (e.g., Yoganarasimhan, 2020) and divided our ground-truth set of images into training, validation, and testing. We assigned 80% of the data to the training+validation set (and use cross-validation) and 20% to the test set. The final accuracy metric of interest was the test RMSE because our ground-truth sentiment and arousal are continuous metrics (refer to Figure 1.3). The expression for the test RMSE is in equation (1.2):

$$RMSE_{test} = \sqrt{\sum_{i=1}^{N_{test}} \frac{\left(\hat{Y}_i - Y_i\right)^2}{N_{test}}}. \quad (1.2)$$

The pool of candidate predictive models included OLS, lasso, elastic net, random forest, SVM, and gradient boosting (regression trees). We implemented machine-learning algorithms with the python package "scikit-learn." Gradient boosting delivered the best test RMSE metric compared to other candidate models. Figure 1.21 plots actuals versus predicted scatterplots for the test set.

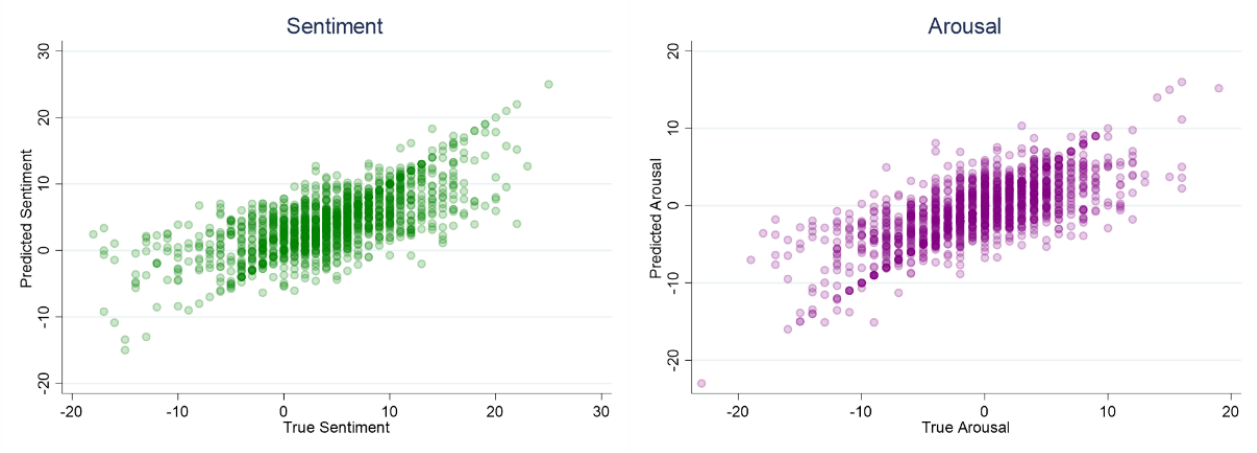


Figure 1.21. Goodness-of-Fit for Test Set, Actual versus Predicted for Sentiment (left) and Arousal (right)

For interpretability, we calculated hit rates for binarized actuals (Y) and predictions ($\hat{Y}$) along various quantiles. For example, for quantile=50%, we median-split Y and $\hat{Y}$ and compute the hit rate from the 2-by-2 confusion matrix. Formally,

$$Hit\_rate_q = \frac{\sum I\left[I\left(Y \geq P_Y^{1-q}\right) = I\left(\hat{Y} \geq P_{\hat{Y}}^{1-q}\right)\right] + \sum I\left[I\left(Y \leq P_Y^{q}\right) = I\left(\hat{Y} \leq P_{\hat{Y}}^{q}\right)\right]}{\sum I\left(Y \geq P_Y^{1-q}\right) + \sum I\left(Y \leq P_Y^{q}\right) + \sum I\left(\hat{Y} \geq P_{\hat{Y}}^{1-q}\right) + \sum I\left(\hat{Y} \leq P_{\hat{Y}}^{q}\right)} \quad (1.3)$$

Here,

- q denotes quantile of interest, $q \in (0,50]$

- $I\left(Y \geq P_Y^{1-q}\right)$ equals 1 if Y exceeds (1-q)-th quantile P_Y^{1-q}, and 0 otherwise
- $I\left(\hat{Y} \geq P_{\hat{Y}}^{1-q}\right)$ equals 1 if $\hat{Y}$ exceeds (1-q)-th quantile $P_{\hat{Y}}^{1-q}$, and 0 otherwise
- $I\left(Y \leq P_Y^{q}\right)$ equals 1 if Y is below q-th quantile P_Y^{q}, and 0 otherwise
- $I\left(\hat{Y} \leq P_{\hat{Y}}^{q}\right)$ equals 1 if $\hat{Y}$ is below q-th quantile $P_{\hat{Y}}^{q}$, and 0 otherwise

Hence, the numerator sums up all cases in which binarized predictions and actuals coincide (number of "successes"), and the denominator sums up all possible cases for a given quantile q (i.e., both "successes" and "failures"). The baseline (no-model) hit rate equals 50% irrespective of the q. Table 1.7 presents hit rates by quantiles for the test set. The accuracy increases as we look at more extreme tails. This could, in part, be explained by the fact that images from the tails (1) were seen by many subjects and hence have a more precise emotion label, and (2) images in the tails represent more extreme emotions, because the overwhelming majority classified these images as either very positive or very negative (analogously, very high arousal or very low arousal). Also, we observe better accuracies for "face present only" and "text present only" subsets of the data. This finding is intuitive because we would expect face and text variables to matter more when we only consider the images that have face/text. Finally, we observe that the model for arousal achieves somewhat greater accuracy than the model for sentiment.

Table 1.7. Hit Rates by Quantiles (Test Set)

	Hit rates by quantiles					
Sentiment model	50%	40%	30%	20%	10%	5%
All data	0.61	0.66	0.71	0.87	0.99	0.99
Face present only	0.65	0.71	0.79	0.91	0.99	1.00
Text present only	0.66	0.73	0.81	0.95	1.00	1.00
Arousal model	50%	40%	30%	20%	10%	5%
All data	0.69	0.73	0.79	0.88	0.97	1.00
Face present only	0.75	0.80	0.85	0.92	0.99	1.00
Text present only	0.78	0.82	0.88	0.95	0.99	1.00

Note. Hit rates by quantiles calculated using equation (1.3). For example, for quantile=50%, the hit rate is computed from the 2-by-2 confusion matrix of median-split Y and $\hat{Y}$.

We next report the relative importance of the emotional modalities. We estimated the full model with all four modalities and recorded the test RMSE, which became a comparison benchmark. We then turned off each of the four modalities one by one, and recorded the drop in RMSE. The size of the relative drop tells us about the relative importance of the excluded modality. We also recorded the RMSE of the null model, which is the intercept-only model. We repeated the procedure for all data, "face present only," and "text present only" subsets of the data. Table 1.8 presents the results. First, as before, accuracy for "face present only" and "text present only" models is better than the accuracy for the entire sample. Second, as before, the arousal model achieves a better fit than the sentiment model. Third, excluding objects/concepts leads to the largest drop in accuracy, whereas excluding visual features leads to the second-largest drop. Face variables matter relatively more for the "face present only" subset but still represent a much lower share of explanatory power than objects/concepts and visual features. Text variables behave similarly to the face variables.

Table 1.8. Relative Importance of the Emotional Modalities Based on the Test RMSE

Panel A. Sentiment model	All data		Face present		Text present	
	RMSE	Dif	RMSE	Dif	RMSE	Dif
Full model	5.34	base	4.92	base	4.47	base
No Visual features	5.57	4%	5.43	10%	4.87	9%
No Objects/Concepts	5.91	11%	5.79	18%	5.55	24%
No Face variables	5.36	0.4%	5.0	2%	4.52	1%
No Text variables	5.36	0.4%	4.98	1%	4.55	2%
Null model	6.37	19.3%	6.6	34%	6.21	39%
Panel B. Arousal model	All data		Face present		Text present	
	RMSE	Dif	RMSE	Dif	RMSE	Dif
Full model	4.83	base	4.26	base	3.83	base
No Visual features	5.03	4%	4.74	11%	4.27	11%
No Objects/Concepts	5.45	13%	5.15	21%	4.88	27%
No Face variables	4.84	0%	4.3	0%	3.85	1%
No Text variables	4.85	0%	4.3	1%	3.88	1%
Null model	5.78	20%	5.76	35%	5.52	44%

Note. "Dif" is the relative increase in the RMSE relative to the full model (base). "Null model" represents the intercept-only model.

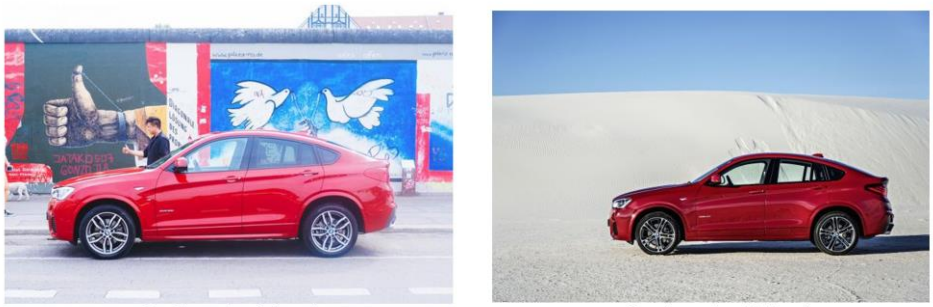
Hence, we uncovered an important insight, namely, that objects/concepts and visual features are more responsible for driving visual sentiment/arousal than facial expressions or text embedded in the image. The result highlights the importance of peripheral features, simple elements of design, for evoking emotions in consumers of visual content. Next, we further examined the emotional impact of individual elements of design.

1.5. Individual elements of design - validation

So far, we have found that objects/concepts and visual features are the primary drivers of visual emotions. Next, for several reasons, we zoomed in on visual features instead of objects/concepts in an attempt to examine the individual impacts of fundamental design elements on visual sentiment and arousal.

First, objects are more numerous and sparser than features (583 objects vs. 114 features). Second, the relationship between features and emotion can be more universal because the objects-emotion link can be contextual / category specific. The concept of "danger" can have different emotional loadings when communicated by a motorbike brand versus a nonprofit. Third, a social media marketer for a given brand might have more freedom over features than over objects. Consider, for instance, two BMW car images in Figure 1.22. The car model and color are the same in both images. However, from the training-set annotations, we know the arousal levels for the two images are drastically different. Because the object is fixed, the difference in emotion is attributable to peripheral visual features.

Hence, by informing social media marketers of what features drive visual sentiment and arousal, we offer them a tool to be used even if the set of high-level objects/concepts they can use is fixed.



Sentiment: +9 (pos)
Arousal: 5 (high)

Sentiment: +11 (pos)
Arousal: -5 (low)

Figure 1.22. Example Images with the Same Object (Red BMW) but Different Emotions.

Note. The numbers come from the training-set annotations.

An advantage of using simple low-level visual features as emotion predictors is the ability to report (and validate) individual elements of design that drive image sentiment/arousal. We used a model that is more transparent than gradient boosting in terms of features' coefficients and allowed

us to measure the statistical significance of the effects. We used the post-Lasso procedure (Belloni & Chernozhukov, 2013), which incorporates, first, running the Lasso model on a full set of variables to eliminate features unrelated to sentiment/arousal, and, second, running OLS on the surviving features to obtain statistical significance. The Lasso regularization parameter λ was chosen via cross-validation separately for the sentiment and arousal models. Table 1.9 presents the results.

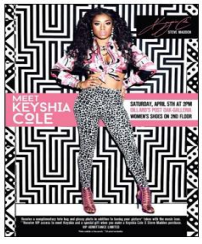
Table 1.9. Important Drivers of Visual Sentiment and Arousal (Post-Lasso OLS)

Visual Feature	Sentiment	Arousal
Dominant color	-1.090**	0^{Lasso}
Hue: green	1.241**	0^{Lasso}
Hue: yellow	-1.279**	0.939**
Hue: orange	1.200**	-1.031***
Hue: red	-1.076***	0.387**
Hue: pink	0^{Lasso}	1.492*
Hue: black-and-white	2.144***	-1.345***
Low brightness	-4.266***	3.679***
High brightness	2.376***	-1.991***
High saturation	0^{Lasso}	3.149***
Concentrated saturation	-1.056**	-0.717**
Amount of edges	5.595***	-2.369
Amount of corners	-442.2*	-259.7
Corner-to-edge ratio	-21.83***	39.02***
Vertical orientation	-0.00169***	0.00112***
Orientation: 45 degrees	0.00249***	0^{Lasso}
Orientation: 135 degrees	-0.00154**	0^{Lasso}
L1-regularization λ	0.0108	0.0073
Observations	10,735	10,735

Note. *** p<0.01, ** p<0.05, * p<0.1. Estimates are obtained by OLS on predictors that survived the Lasso regularization (10-fold cross-validation). Variables that did not survive Lasso regularization for either sentiment or arousal are marked 0^{Lasso}.

Next, we validated the impacts of the most important drivers from Table 1.9 in the lab. We conducted a lab study (N=300) in which half of the subjects were shown an original image sampled from a brand's Twitter account, and the other half were shown a modified version of the same image. We modified the image by changing a particular visual feature (element of design) that was found to be associated with the visual sentiment/arousal in the training exercise. For example, for image A in Table 1.10, we decreased the amount of pink and saturation. For image B, we decreased the color variety. For image C, we reduced the number of corners in the background. We altered the dominant color in image D and rotated image E to change the line orientation.

Table 1.10. Examples of Original versus Modified Images from Lab Validation of Drivers of Visual Sentiment/Arousal

A.	**Saturation + pink**   **Sentiment: positive** 90% vs 39% (p<0.001) **Arousal: high** 58% vs 25% (p<0.001)		Image on the right was modified to have a reduced level of pink hue and lower saturation.
B.	**Color variety**   **Sentiment: positive** 95% vs 66% (p<0.001) **Arousal: high** 38% vs 36% (p=0.59)		Image on the right was modified to have a lower color variety. Hues in the original image are more distinct.
C.	**Many corners**   **Sentiment: positive** 61% vs 60% (p=0.75) **Arousal: high** 84% vs 78% (p=0.028)		Image on the right was modified to have fewer corners in the background.

D.

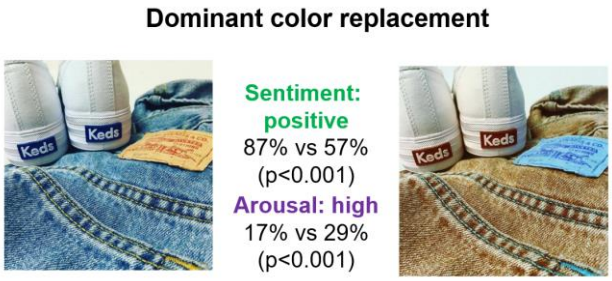


Image on the right was modified to have the orange dominant hue instead of the blue.

E

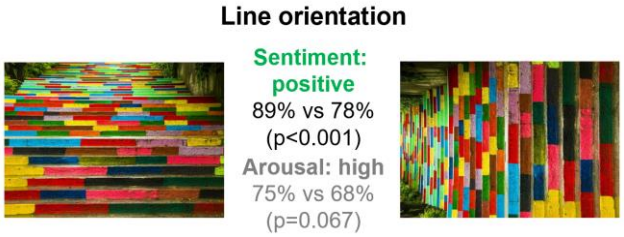


Image on the right was rotated to exhibit vertical rather than horizontal orientation.

Note. Each pair was rated by 300 subject lab participants. Chi-squared test p-values are reported.

We found significant differences in subjects' sentiment/arousal scores for modified (vs. original) images, consistent with our model predictions. The visual features we identified (e.g., pink, saturation, color variety, corners) drive human perceptions of sentiment and arousal of an image. We conducted this validation exercise for 26 original images and for 20 features, and found significant differences in 77% of the cases. This experiment validates the findings that the elements of design reported in Table 1.9 indeed significantly affect the levels of visual sentiment and arousal. We believe these insights are important for visual designers and social media marketers.

1.6. Discussion

In this research, we proposed and validated a tool to predict an emotion of a digital image given purely pixel-level information from the image. First, we collected a novel ground-truth annotated set of images on the two primary dimensions of emotions: sentiment and arousal. We had 1,292

subjects annotate 13,386 original brand-generated images from Twitter. The images came from 637 brands from 11 categories and were posted during 2011-2017.

Second, we extracted visual characteristics from the images that correspond to four emotion modalities: (1) elements of design (low-level visual features) such as color, texture, shape, lines, curves, corners, edges, and orientation; (2) high-level visual objects/concepts (e.g., adventure, action, leisure, danger, etc.); (3) human facial expressions; and (4) text embedded in the image. We used computer vision and deep-learning algorithms to extract the variables for all four modalities. We then used the variables to predict sentiment and arousal in our training set of images. The resulting predictive model is highly accurate. We evaluated the relative importance of the modalities and found that high-level visual objects/concepts and elements of design (low-level visual features) are much stronger drivers of visual emotions than facial expressions and the text.

Third, we reported and validated individual elements of design - drivers of visual emotion. For example, from the predictive model, we found that color variety, non-smooth texture, higher amounts of green, and orange hues in an image, higher brightness, and horizontal orientation are associated with more positive sentiment of the image. Many corners, red hues, and pink hues are associated with high arousal. We validated each of the independent drivers in the lab by modifying original firm-generated Tweet images to have less (or more) of a particular visual feature. For 77% of the image pairs, we found statistically significant differences in the scores assigned by human subjects, and these differences agreed with model predictions.

Further directions for the research include examining the heterogeneity in the viewer's perceptions of visual emotions. First, subjects' annotation patterns can be linked to subjects' characteristics (e.g., demographics). Understanding how the target audience perceives the

emotionality of visual content is important for a social media marketer. Second, heterogeneity of individual drivers of emotions should be examined. So far, we have obtained the average effects (e.g., color variety drives visual sentiment) but have not considered potential heterogeneity. Third, investigating interactions between emotion modalities (e.g., red+child) in their impact on visual emotions more transparently, and potentially conducting A/B tests similar to ones in section 1.5, but for the interactions, might be fruitful.

Essay 2. The Role of Images and Words: Understanding Engagement with Firm-Created Social Media Content

2.1. Theory of Persuasion and the Role of Text and Imagery

The study of persuasion and resistance to persuasion has a long tradition in psychology and consumer behavior. The persuasion knowledge theory states that once the targets are aware of a persuasion attempt, the targets/receivers change their attitude (and behavior) toward the message coming from the agent/sender. Regarding FGC on social media, the degree to which users activate defensive processing may affect the likelihood of acceptance, "liking," and content sharing by consumers. Therefore, understanding defensive processing patterns of users as they relate to both text and visual elements of social media content is crucial for brand owners.

Friestad and Wright (1994) suggest a framework of how people form and use persuasion knowledge to cope with persuasion attempts. The development of persuasion knowledge is dynamic within an individual and is subject to cultural and generational effects. Individuals are able to learn from a particular marketing tactic and to develop the persuasion knowledge of this tactic over time. An individual's general persuasion knowledge tends to increase with age. Regarding a specific marketing environment (e.g., certain social media network such as Twitter), this implies a user's response to a given persuasion tactic will change over time as the user develops persuasion knowledge of this tactic.

Campbell and Kirmani (2000) further expanded on the persuasion knowledge model by examining the roles of the target's cognitive capacity and of the accessibility of the agent's motives. When the agent's motives are easily accessible, both (cognitively) busy and non-busy targets use the persuasion knowledge to scrutinize the agent's sincerity. However, when the agent's

motives are not easily accessible, cognitively busy targets are significantly less likely to use persuasion knowledge to evaluate the agent.

Another popular persuasion model, the elaboration likelihood model (ELM), developed by Petty and Cacioppo (1986), suggests persuasion usually occurs along two routes: central and peripheral. Under the central route, the target shows a high level of elaboration by scrutinizing arguments in the persuasive message. Under the peripheral route, the target uses cues that are not directly related to the message meaning or content to decide on his or her attitude toward the message. The peripheral route represents a low level of elaboration by the target. A classic example of a peripheral cue is source credibility. Importantly, attitudes created via central versus peripheral routes are different in nature. Attitudes formed through the central route are strong and persistent and require counterarguments to weaken. Attitudes formed through peripheral cues create fleeting and unstable attitudes that are not likely to be strongly protected by the target.

Both the persuasion knowledge model and the ELM are relevant in our research setting. Because the rise of visual content is a relatively recent phenomenon in social media (on Twitter, since 2011) compared to text-based content, the persuasion knowledge model suggests users are better able to resist text-based persuasive tactics used by marketers than visual-based tactics. That is, because consumers have been exposed to text-based marketing on social media for a longer time, they could have developed stronger defensive mechanisms against text-based marketing persuasion tactics. Further, the ELM suggests messages containing images along with text (vs. text-only messages) are likely to activate peripheral cues and, as such, have a greater chance of influencing user attitudes, particularly in low-user-involvement contexts (Mitchell, 1986; Miniard et al., 1991). As such, because consumers have had less exposure and experience with imagery-based marketing tactics on social media, and because imagery tends to activate peripheral cues,

imagery may be more persuasive. We derived the following hypotheses stemming from these arguments:

H1. *Over time, consumers have developed defensive processing of text-based persuasive messages posted by brands in social media, rendering text-based marketing messages less effective.*

H2. *Consumers are better able to resist persuasion through text rather than imagery in firm-generated social media communications. Particularly, high motivation-to-act text will face significantly more resistance than high motivation-to-act imagery.*

H3. *Image-based strategies allow marketers to achieve higher maximum potential impact on engagement than text-based strategies.*

2.2 Data and model

In this section, we examine the relationship between the emotional content of text and images in a social media message and consumers' engagement with this message. We analyzed Twitter timelines of 637 brands from 11 categories (both commercial and charity/nonprofit), starting from 2008 (modal brand Twitter account was created in 2009). The brand sample is from the Young and Rubicam Brand Asset Valuator (BAV, Lovett et al., 2014), which is a large-scale representative survey of brand perceptions in the US. We manually linked brand names from the list to their Twitter handles and verified each account. Figure 2.1 graphs the overall volume of Tweets in our sample by year and breaks it down by category.

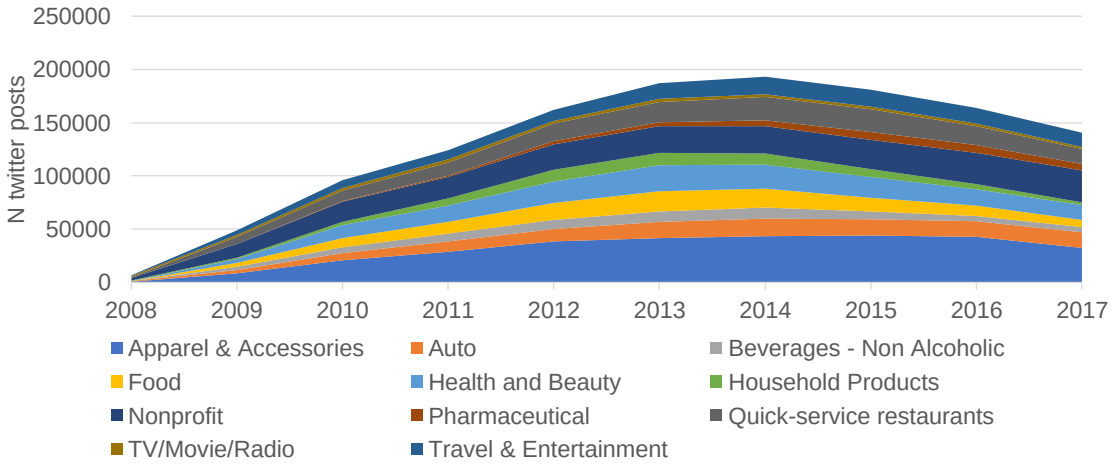


Figure 2.1. Volume of Tweets by Brand Categories in the Empirical Dataset

Twitter API allows us to obtain the 3,200 most recent Tweets for a given account. To bypass this limitation and download older Tweets, we undertook a two-step procedure. First, we used the Selenium automated browser along with Twitter advanced user search (which allows the specification of "from" and "to" dates when searching for tweets from a particular account) to obtain tweet IDs for each account. Next, we used Twitter API to obtain information on all Tweet IDs in our list. We obtained data on 1,345,473 Tweets, 449,209 of which have visual content. For each Tweet, we obtained the number of Retweets, timestamp, brand/account name, Tweet text, and URL of image/video (if present). Tweets in our sample do not include Retweets (e.g., brand A Retweeting a Tweet of brand B) or replies to other users—we focused purely on original Tweets that a brand posts in its timeline for all followers to see.

As a dependent variable, we used the cumulative number of Retweets for a given Tweet as the date of data acquisition. We used a static cumulative number of Retweets, because 90.4% of all Retweets happen within 24 hours since the Tweet upload. To validate this, we undertook a side data-scraping exercise (refer to Appendix A for details). Because Retweets "flatten-out" within 24 hours for the vast majority of Tweets, we assumed our Y and X variables align sufficiently in time.

Table 2.1 presents summary statistics for Tweets in our sample. Because the distribution of raw Retweets is highly skewed, we log-transformed the Retweets for the empirical analysis.

Table 2.1. Summary Statistics of the Dependent Variable

Variables	N	Mean	Median	SD	Min	Max
Retweets	1,345,473	29.5	4	312	0	122955
log(1+Retweets)	1,345,473	1.8	1.6	1.53	0	11.72
Retweets (img present)	449,209	54	10	488	0	122955

Table 2.1 suggests Tweets with images score higher on engagement. We estimated a more precise effect of image presence using a model with "brand — week-of-the-year" fixed effects. Our final goal, however, was not to estimate the effect of image presence, but rather the impact of image sentiment and arousal on engagement, and compare it with the impact of text sentiment and arousal on engagement. We applied the procedure of visual sentiment/arousal scoring (described in section 1) to the list of Tweet image/video URLs (see Figure 1.5 for the examples of images). To be able to assess the relative impact of image versus text on engagement, we needed to obtain measures of sentiment and arousal for a Tweet text.

2.2.1. Quantifying sentiment and arousal of Tweet text

We quantified the emotional content of text along two dimensions: sentiment and arousal. To quantify the sentiment of text in a corporate Tweet, we used VADER, optimized specifically for Twitter posts (Hutto & Gilbert, 2014). The algorithm uses annotated vocabulary to assess Tweet-text valence on a scale from -1 to 1. It incorporates information from emoticons and takes into account the intensity of a particular sentiment (e.g., ":)))" will have a more positive Sentiment than ":)"). Examples of Tweets with different sentiment levels follow:

[-] "Mesothelioma is a rare but fatal form of cancer that is often difficult to diagnose." vs.

[+] “Waffles make people happy. Like, really happy. Like, best way to start your day happy.”

To compute text arousal of brand tweets, we tracked usage of specific characters and punctuation in Tweet text (Barbosa & Feng, 2010), as well as usage of “call-to-action” words that are used to prompt a certain action from the audience. Specifically, a high usage of exclamation and question marks, a high proportion of capital letters, and usage of “call-to-action” words[2] would render a Tweet high-arousal. Examples of Tweets with different arousal levels follow:

[High] “ITS NATIONAL CHOCOLATE DAY! DROP EVERYTHING AND EAT CHOCOLATE!“ vs.

[Low] “Beer at the end of a long day makes everything better.”

Having obtained four main independent variables of interest (text and image sentiment and arousal), we formulated an empirical model for engagement (Retweeting) and provide below a detailed description of controls.

2.2.2. Empirical Model

We employed regression on a Tweet level with “brand — week-of-the-year” fixed effects. For Tweet i posted by brand j,

$$Log\left(1 + Retweets_{ij}\right) = c + \sum_{j=1}^{N_{brands}-1} \sum_{k=1}^{51} brand\_FE_j \times week\_of\_the\_year_{k(i,j)} + TextStrategy_{ij} \times \left[1 + \sum_{k=2009}^{2017} year_{k(i,j)}\right] + ImgStrategy_{ij} \times \left[1 + \sum_{k=2009}^{2017} year_{k(i,j)}\right] + TimeControls_{ij} + TextControls_{ij} + ImgControls_{ij} + e_{ij.} \quad \text{(Eq 2.1)}$$

[2] The list includes ask, come, try, share, buy, save, grab, invite, enjoy, sip, choose, taste, smell, celebrate, explore, go, fly, savor, submit, pick, play, join, participate, download, donate, find, obtain, and visit.

- $Log(1 + Retweets_{ij})$ is a log transformation of the number of retweets of tweet i posted by brand j (cumulative as per date of data acquisition[3]). Log transformation of the engagement metric is a frequent choice in many studies examining the impact of content characteristics (e.g., Lee, Hosanagar, & Nair, 2018; Stephen, Sciandra, & Inman, 2015)

- $brand\_FE_j$ is the brand fixed effect.

- $week\_of\_the\_year_{k(i,j)}$ is the week-of-the-year effect (Tweet belonging to one of the 52 weeks of the year). Interactions with brand FE-s give 51 dummies for each brand. Fixed effects at this level allowed us to estimate brand-specific seasonality in both engagement and content strategies, thus helping us rule out potential tendencies of brands to post particular content at a particular time of the year (e.g., Starbucks posting red cups around Christmas). All our effects were identified from deviations from "brand — week-of-the-year" means. Because our fixed effects are numerous (30K+) and have high-granularity, they alone explain approximately 50% of the variation in log(1+Retweets).

- $TextStrategy_{ij}$ includes four possible combinations based on the median split (binary) sentiment and arousal of tweet text. "Negative sentiment, low arousal" serves as the base text strategy.

- $ImgStrategy_{ij}$ includes five possible image strategies: "no image," "negative, low-arousal image," "negative, high-arousal image," "positive, low-arousal image," and "positive, high-arousal image."

- $TimeControls_{ij}$ include the following:

[3] We found that 90.4% of Retweets occurred, on average, within the first 24 hours after a Tweet upload. Hence, our main outcome metric (cumulative number of Retweets) is largely attributable to first 24 hours after the Tweet upload (we observe upload day), which in turn suggests our outcome metric and text-/image-strategy variables are largely aligned in time, and the diffusion-in-time effect for an average Tweet is minimal.

- $month\_year_{ij}$ – calendar monthly trend common to all brands in the sample
- $hour_{ij}$ - upload-hour effect (23 dummies),
- $day\_of\_week_{ij}$ - upload-weekday effect (6 dummies).

We interacted text- and image-strategy variables with year dummies to obtain the time paths of the effects on engagement. Next, we describe Tweet-text and -image controls in detail.

2.2.3. Tweet-text controls

$TextControls_{ij}$ include the length of the Tweet (and its square), the number of hashtags (and its square), the presence of URL in the Tweet, the presence of the promotional keywords "sale," "discount," and "% off." We also controlled for **Tweet-text meaning** by using pre-trained *word vectors (embeddings)* (Mikolov et al., 2013; Timoshenko & Hauser, 2019). Word embeddings are products of a neural network that is trained on a large corpus of text. The network considers the *context* in which a word occurs, hence representing the semantic meaning attached to the word. Using vector operations, meanings could be added and subtracted from words (e.g., *Paris – France + Italy = Rome*). Equivalently, words with similar semantic meaning would locate closer to each other in the vector space. We used word embeddings that were already pre-trained on a corpus of 400 million Twitter microposts (Godin et al., 2015). Each word in a Tweet text is scored on 400-dimensional word vectors. To aggregate embeddings on a weet level, we averaged vector scores across the words in a Tweet. Next, to reduce dimensionality of word vectors, we first calculated 200 principal components (PCs) that capture ~80% of the variance in word embeddings. As a result, we added 200 more control variables that proxy for Tweet-text meaning. Importantly, the addition of word-vector PCs to the right-hand side of the equation increased adjusted R-squared

by 7% (over and above the fixed effects). Also, 168 out of 200 word-vector PCs had a significant ($p<0.01$) effect on engagement.

2.2.4. Tweet Image Controls

First, we controlled for the presence of **high-level visual objects (concepts/tags)** in a Tweet image (e.g., Figure 2.2). We used a deep-learning tool (Clarifai.com) to extract 20 objects from each image in our sample (5832 unique objects in total). Each extracted concept also has a confidence score. We retained objects that have a confidence score of at least 97% and appear at least in 100 times in our data, leaving us with 1,042 unique objects. We created a dummy variable for each object and added them to the right-hand side of the engagement equation. Out of 1,042 objects, 331 are significant at 1% or less.

Object	Certainty	Object	Certainty
wind	100%	generator	92%
sky	98%	ecology	92%
turbine	98%	windmill	89%
nature	95%	environment	89%
invention	95%	sustainability	87%
electricity	94%	outdoors	87%
cloud	94%	fair weather	87%
power	94%	wind turbine	87%
technology	92%	energy	86%
grinder	92%	industry	84%

Figure 2.2. Example of Tweet Image and List of High-level Visual Objects Extracted Using Clarifai.com

Second, we controlled for the presence of **embedded text in an image**. We used Pytesseract, optical recognition software (OCR), to extract strings of text from image files. Figure 2.3 provides examples of extraction. We found that ~25% of Tweet images have some identifiable text in them. If OCR text is present, we scored its sentiment and arousal using the same procedures described in section 2.2.1. We also extracted length, the presence of URLs, hashtags, and promo keywords. We added these variables, along with the dummy for OCR text presence, as additional controls to the right-hand side. We found that OCR-text sentiment and arousal effects on engagement are very similar to the effects of the main Tweet-body text.

Original image:

Text extracted:

```
...: text=pytesseract.pytesseract.image_to_string(img)
...: print(text)
...:
a special treat for our twitter followers!
SURPRISE SALE

GIFTS $99 & UNDER

(all wrapped up and ready to go!)

PLUS FREE GROUND SHIPPING
TO THE U.S. & CANADA WHEN YOU SPEND $75 OR MORE



ALL SALES FINAL. ENDS DECEMBER ISTH AT 11:599M PT.

In [140]:
```

Figure 2.3. Application of OCR (Optical Character Recognition Software) to Extract Embedded Text in Tweet Messages

Third, we assessed the **color distance between Tweet image and brand logo**. Certain brands tend to post imagery that resembles their logo / trade dress. Without explicitly controlling for this tendency, we could misattribute its effect to effects of image sentiment/arousal. We collected logo image files from Twitter accounts of brands and extract a vector of hues (20 bins) for each logo using the same technique as described in section 1.3.1. Then, for each Tweet image, we computed the Euclidean distance from the associated logo in a 20-dimensional space of hues. Figure 2.4 provides examples of images and logo color distance.

Logo	img1	img2	img3	img4
Distance=0	0.04	0.26	0.64	0.95

Figure 2.4. Examples of Same-Brand Images and Associated Metric of Logo Color Distance

Figure 2.5 summarizes all types of variables that we extracted from a brand Tweet. Overall, we found our full model explains approximately 70% of the variation in log(1+Retweets), with roughly 50% coming from fixed effects and 20% coming from time controls, text, and image variables.

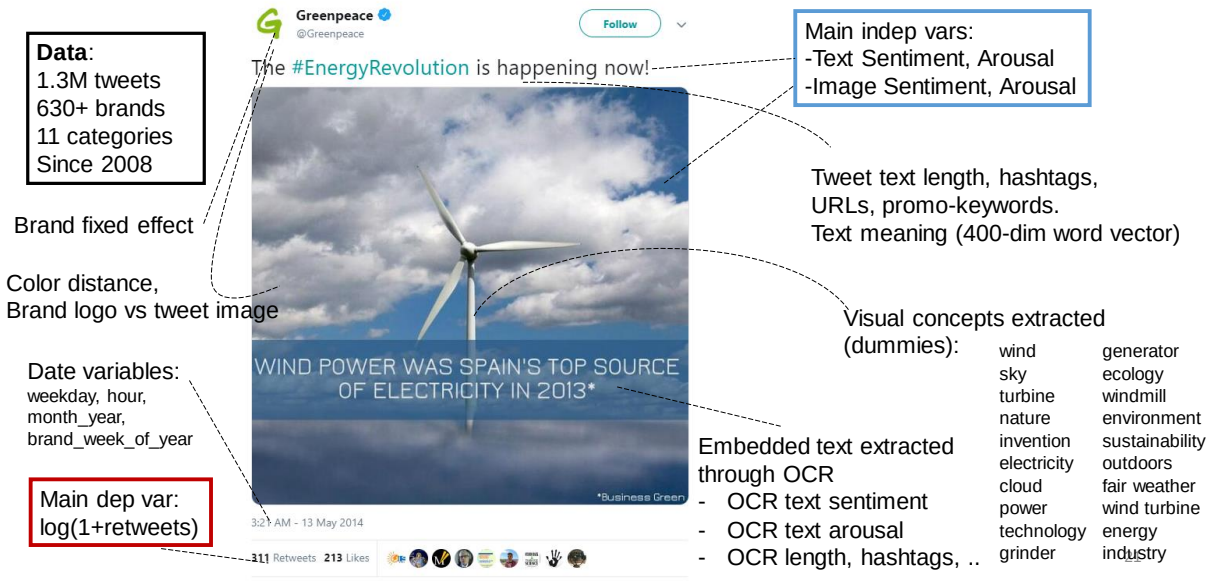


Figure 2.5. Summary of Variables/Controls Involved in the Empirical Analysis

2.3. Results

Examining the effects of text sentiment and arousal over time, we found high-arousal (high motivation-to-act) text and positive-sentiment text were effective in engaging customers up until 2011. After 2011, they started to have a negative impact on engagement (Figure 2.6). This finding is interesting and consistent with the persuasion knowledge model (Friestad & Wright, 1994; Campbell & Kirmani, 2000), which suggests user response to a persuasion tactic will diminish over time as users develop persuasion knowledge of this tactic. We also found text sentiment is associated with lower engagement, particularly for Tweets that also employ low-arousal text.

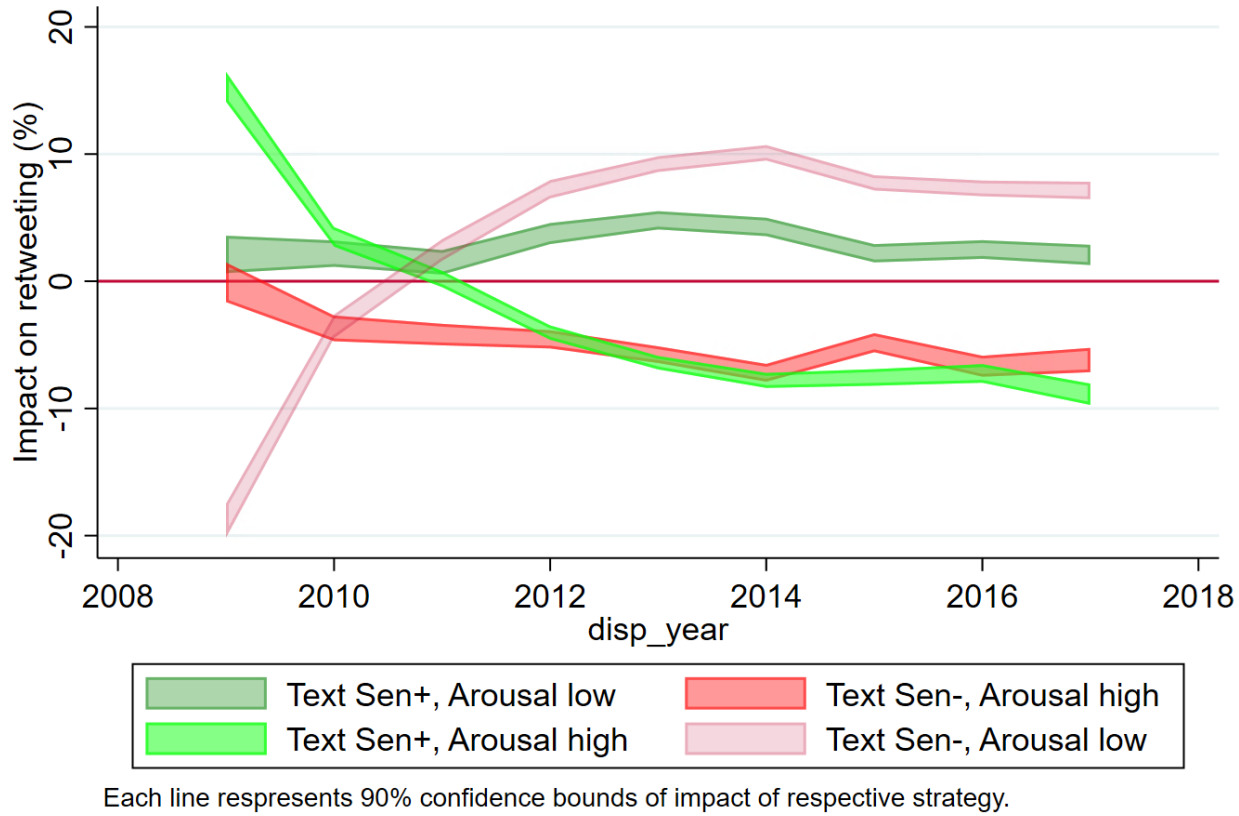


Figure 2.6. Effect of Tweet-Text Strategies on Retweeting.

Note. Model without intercept allowed to obtain all four text-strategy paths.

We detected a similar decline in the impact of promotional keywords in Tweet text (Figure 2.7). We also linked this finding to the predictions of the persuasion knowledge model.

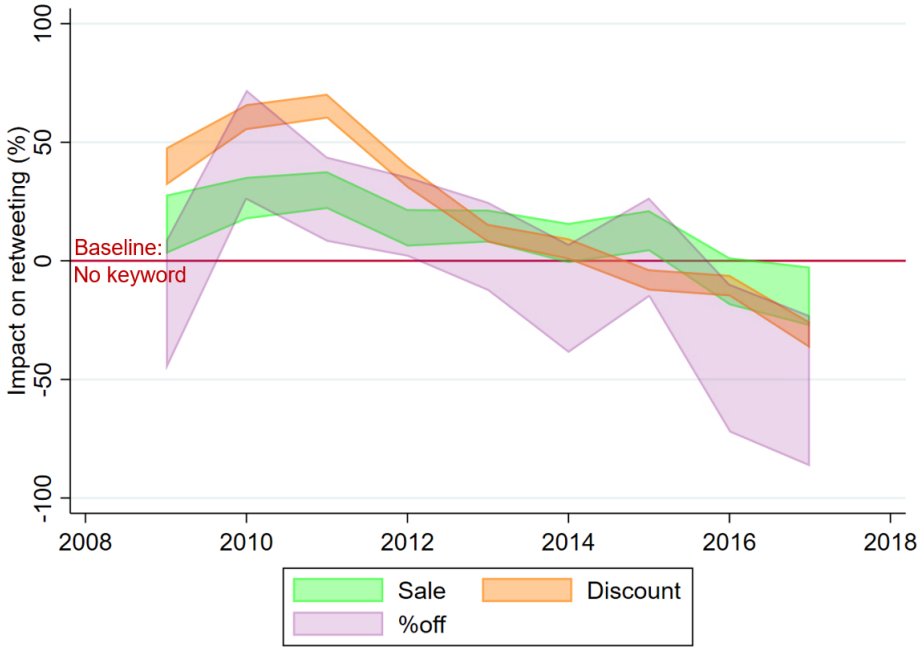


Figure 2.7. Effect of the Presence of Promotional Keywords in Tweet Text on Retweeting

Interestingly, we did not detect a similar decline in the effectiveness of high-arousal (high motivation-to-act) imagery (Figure 2.8). We found high-arousal and negative-sentiment imagery is a dominant strategy (at least weakly, and sometimes strictly). Comparing it with the inferiority of high-arousal text strategy, we suggest consumers might be better able to resist persuasion through text rather than imagery. These results suggest *marketers should* ***decrease activation through text and instead leverage imagery to activate consumers*** *in their social media communications.*

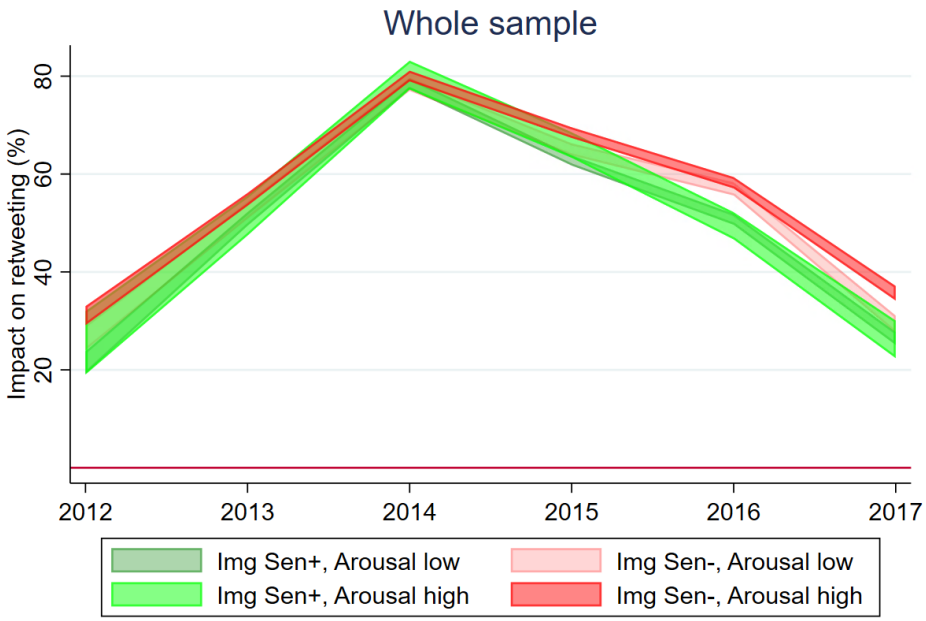


Figure 2.8. Effect of Image Strategies on Retweeting

We also uncovered significant heterogeneity in image effects by repeating analysis by sector. Figure 2.9 highlights results for three categories: charities/nonprofits, quick-service restaurants (QSR), and health and beauty. We found that, for charities/nonprofits, negative visual sentiment is a dominant strategy, and image arousal has a second-order effect. The optimal visual strategy for QSR is positive, high-arousal imagery, contrary to the whole-sample results. For health and beauty, we found image arousal has a first-order effect, with highly activating imagery being a dominant strategy.

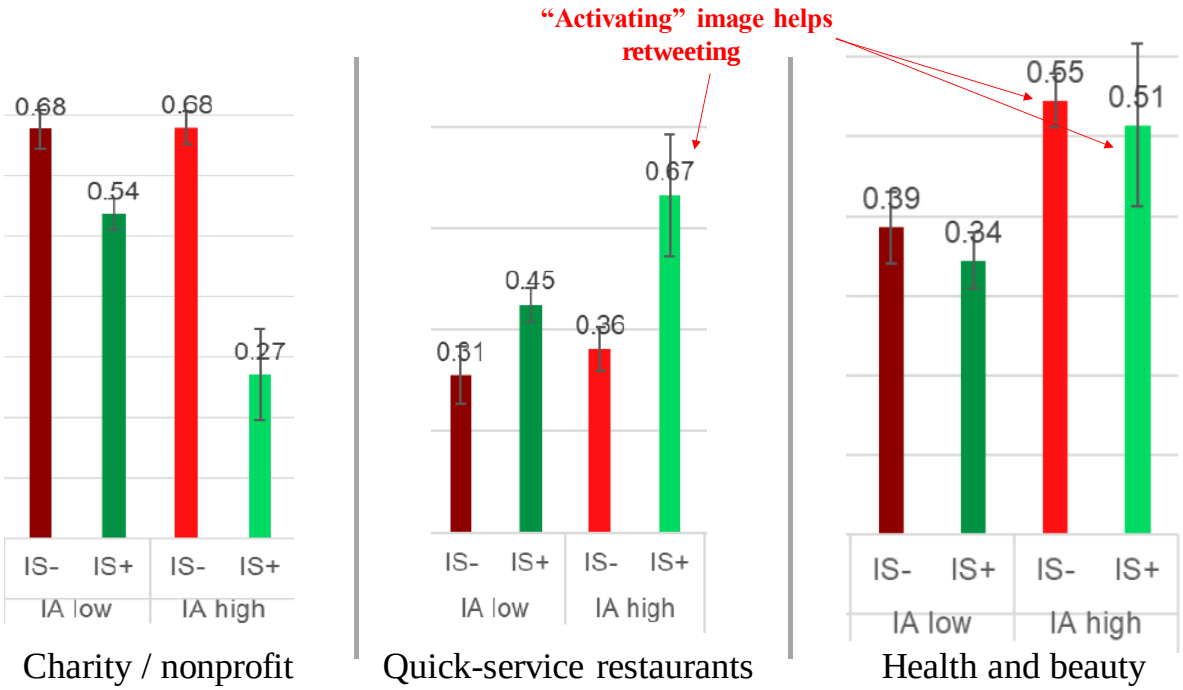


Figure 2.9. Heterogeneity of Visual-Strategy Impact over Different Categories

Note. IS = Image Sentiment, IA = Image Arousal/Activation.

Based on the industry-specific strategy-impact estimates, we can identify dominant text and visual strategies by sector (Table 2.2). We also calculated total potential gains from employing the dominant strategy, by computing the difference between the best-performing and worst-performing strategies (provided they are statistically significantly different).

Table 2.2. Dominant Strategies by Sector

	% of the sample	Median retweets	Dominant text strategy	Dominant Img strategy	Max impact of text	Max impact of Img at dominant text
Whole sample	100%	4	TS-, TA low	IS -, IA high	11%	13%
Apparel & Accessories	22%	4	TS+, TA low	IS -, IA any	10%	6%
Auto	8%	13	TS+, TA low	IS any, IA low	11%	8%
Beverages - Alcoholic	3%	4	TS-, TA low	none	21%	0%
Beverages - Non Alcoholic	5%	2	TS-, TA low	IS any, IA high	19%	11%
Food	9%	1	TS-, TA low	IS -, IA any	8%	17%
Health and Beauty	11%	3	TS any, TA low	IS -, IA high	13%	22%
Household Products	4%	1	TS+, TA high	IS any, IA high	7%	11%
Nonprofit	15%	9	TS-, TA low	IS -, IA any	18%	41%
Pharmaceutical	3%	5	TS any, TA low	IS -, IA any	15%	38%
QSR	10%	4	TS-, TA low	IS +, IA high	31%	36%
Travel & Entertainment	8%	3	TS-, TA low	IS any, IA low	6%	16%
Weighted average across sectors					**13.9%**	**19.1%**

Note: "none" means visual-strategy effects were statistically indistinguishable from one another.

When averaging across sectors, we found the image strategy suggests the potential of increasing engagement within 19%, and text, within 14%. A higher potential maximum impact of image-based strategies supports H3 and could potentially be explained by ELM (Petty & Cacioppo, 1986): Visuals might activate additional peripheral cues for persuasion and hence have a higher chance of influencing users, particularly in low-user-involvement contexts (Mitchell, 1986; Miniard et al., 1991).

We also examined the potential interplay in image and text effects. We augmented equation (2.1) to include interactions between text strategies and image strategies, which resulted in 16 coefficients (four strategies for text and four for image). Figure 2.10 presents the results. For instance, fix text sentiment at negative, low-arousal (dark-green cells). Then, depending on the image emotion, the bottom-line impact on engagement would range from 6.1% to 8.4% (albeit not statistically significant). The deltas hence represent the maximum possible interplay effect for

adjusting image emotion for a fixed text emotion. Note that most of the deltas are small or insignificant, with the biggest delta equal to 2.8%. Compared to the main effects of text and image emotions (e.g., Figures 2.6 and 2.9), the text-image interplay effects appear to be of second-order magnitude. We conclude that text and image drive engagement mostly independently rather than jointly.

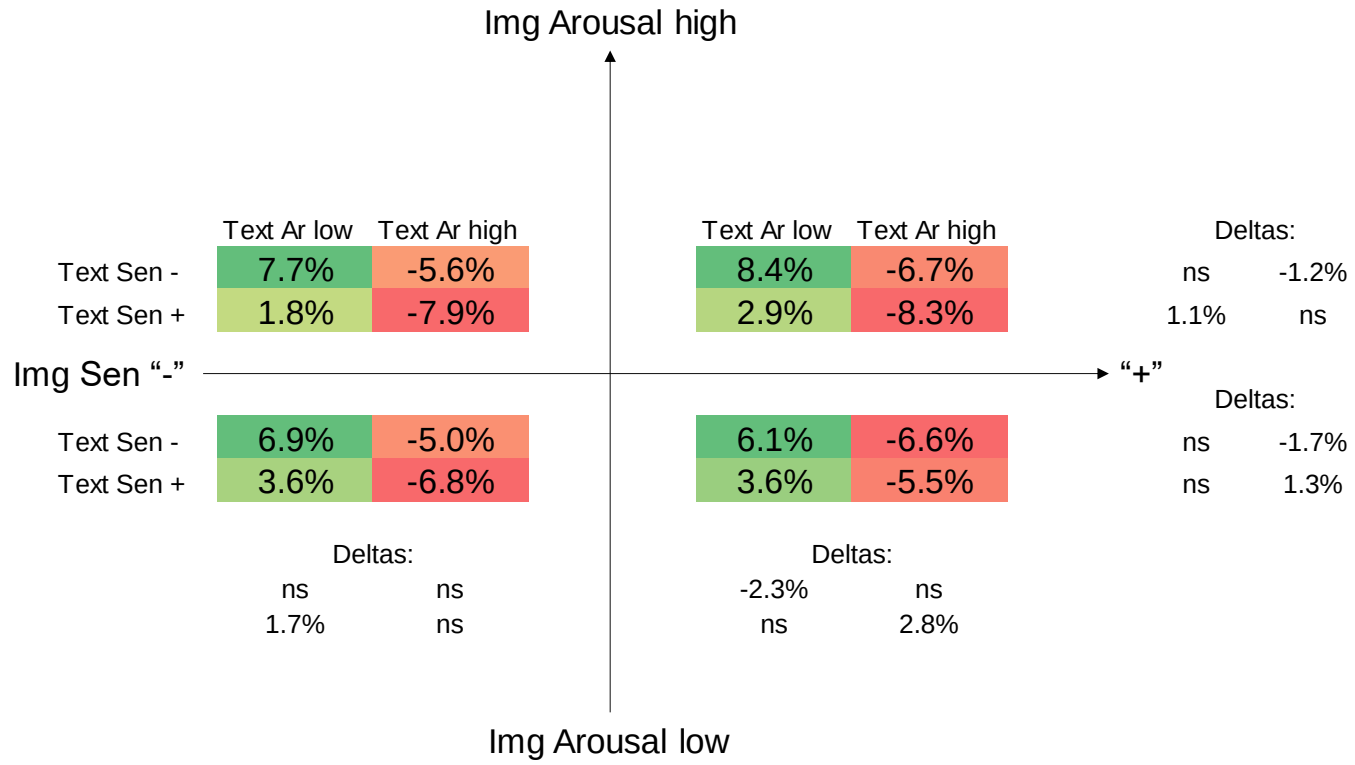


Figure 2.10. Interplay in Image and Text Effects.

Note. Deltas denote maximum possible impact of adjusting image emotion for a fixed text emotion. Delta is "ns" if the interplay effect is not significant at the 5% level.

2.4 Discussion

We undertook an empirical analysis of engagement (Retweeting) with 1.3 million Tweets from 630+ brands from 11 categories, posted since 2008. We used a visual-emotion scoring tool to create variables of visual sentiment and arousal for Tweet images. For Tweet-text sentiment, we

used an existing tool (VADER) optimized specifically for Twitter posts. Tweet text is higher on arousal/activation scores if it uses exclamation marks, question marks, a high proportion of uppercase letters, and/or "call-to-action" verbs (e.g., try, share, buy, etc.). We examined the impact of image/text strategies on Retweeting using a "brand — week-of-the-year" fixed-effects regression with multiple controls such as Tweet-text meaning, high-level visual objects, text embedded in images, logo color distance, and so on. The full model explains 70% of the variation in Retweeting. We found the effectiveness of the high-arousal/motivation-to-act text strategy has decreased over time, with the impact switching from positive to negative around 2011. We suggest this finding could be explained by Friestad and Wright's (1994) persuasion knowledge model: Over time, consumers have become significantly more resistant to high-activation text. Interestingly, we did not find the same pattern for high-activation imagery, with the high-arousal image strategy staying dominant for images that also exhibit negative visual sentiment. These results suggest *marketers should* ***decrease activation through text and instead leverage imagery to activate consumers*** *in their social media communications*. Further, we found significant differences in dominant visual strategies by sector, with, for example, negative-sentiment images being dominant for charities/nonprofits, and positive-high-arousal images being dominant for quick-service restaurants. We estimate the industry-specific dominant image strategy has the potential to increase engagement by up to 19%, whereas the text strategy has the potential to increase engagement by up to 14%.

Further efforts should focus on a more granular examination of the heterogeneity of the effects. We did not zoom in on the categories to understand brand-specific heterogeneity. Given that a single brand may upload thousands of Tweets, the data should be rich enough to examine brand-level heterogeneity. Another direction for research is further investigation of the

psychological mechanism behind the effects. We suggest the interpretation of the effects based on persuasion knowledge and elaboration likelihood models, but our findings do not causally prove this mechanism. Finally, we considered Retweets as an engagement metric, but looking at alternative metrics such as likes, comments, and follower base might be interesting.

Essay 3. Effects of Face and Gaze in a Product Image on Browsing and Ordering

3.1 Introduction

Human faces are powerful drivers of perceptions and judgments, and eyes are the most attended region of a face (Sajjacholapunt & Ball, 2014). Yet, no studies exist on the effects of the face and gaze of models in the product images on outcomes such as product clicks, orders, and returns. This current research applies deep-learning algorithms for face detection and gaze-following in the context of 57,088 apparel products from 22 categories from one of the largest Chinese e-commerce websites. We document interesting and important implications of the face and gaze of models in product images.

First, we show that consumers more often browse product images that include the model's face. However, higher prominence of the face leads to less browsing but more ordering of the product. With faces currently present in only 30% of apparel pictures in our setting, marketers must understand the tradeoffs involving face usage and face prominence.

Second, we document a negative impact of the "direct" gaze (i.e., the model's gaze directed toward the viewer) on both browsing and ordering. The finding is consistent with previous behavioral studies that looked at metrics such as dwell time on the product/brand elements of the ad, memorization, and recall. This study is the first to demonstrate the effects of the "direct" gaze in a large-scale empirical setting. Given that the "direct" gaze is used in 30% of the images that have a face, this insight is importantfor practitioners.

Third, this study is the first to consider subtypes of the "averted" gaze (i.e., the model's gaze that is not directed twoard the viewer), which are "sideways" gaze and "downwards" gaze. Interestingly, "downwards" gazes result in fewer browses and orders than "sideways" (particularly, "to-the-right") gazes. We suggest a potential explanation for underperformance of

the “downwards” gaze stemming from potential associations with shame and embarrassment (Clifford & Palmer, 2018).

Taken together, our results suggest the following implications for product photo designers. Overall, featuring a model’s face in a product image is beneficial. Face prominence should be higher to achieve more orders and a better conversion rate, but lower to achieve more product browses. Next, a model’s gaze should be neither direct nor downwards. A sideways gaze is a solution, and a right-directed gaze works somewhat better than a left-directed gaze.

Next, we review existing literature on the face and gaze effects, describe our empirical setting, describe the deep-learning tools for face and gaze detection and extraction, and report the empirical findings.

3.2 Face and gaze effects – theory

3.2.1 Existing research on faces

Human faces are powerful drivers of perceptions and judgments. Discussion of physiognomy and the tendency of people to make inferences from faces dates back to at least the 18th century (Willis & Todorov, 2006). People are quick to make judgments based on facial features. Judging a stranger’s face on dimensions of attractiveness, likeability, trustworthiness, competence, and aggressiveness takes less than 100ms (Willis & Todorov, 2006). Longer exposures make people even more confident in their initial judgments.

Todorov et al. (2005) showed that voters’ perception of a candidate’s face portraying competence is associated with a higher chance of this candidate winning in the US Congressional elections. Zebrowitz and McDonald (1991) found that baby-faced individuals are more likely to receive milder sentences in court than mature-faced individuals. The same face will also be judged differently by different people depending on the perceiver’s personality traits, demographics,

disorders, or even transient goals (Zebrowitz, 2006). This research demonstrates how people make rapid judgments based solely on facial features and how these judgments shape important real-world outcomes.

In marketing, researchers have studied faces in the context of anthropomorphism (e.g., Aggarwal & McGill, 2007), celebrity/spokesperson perceptions (e.g., Gorn et al., 2008), and models' faces in print ads (Xiao & Ding, 2014). Maeng and Aggarwal (2018) examine consumer preference for products (e.g., car front, watches) that exhibit higher/lower width-to-height ratios. In psychology, a higher width-to-height ratio of the human face is linked to perceptions of dominance. Maeng and Aggarwal (2018) found consumers prefer products with higher width-to-height ratios because such purchases allow them to signal their own dominance/status.

Wang et al. (2017) examined consumer perceptions of a marketer's smile width (slight vs. broad). Findings show marketers with a broader smile are perceived as higher on warmth but lower on competence. Using data from Kickstarter, the authors demonstrated the negative correlation between an entrepreneur's smile width and the average contribution per backer, which could be explained by the backers perceiving the entrepreneur as less competent.

Gorn et al. (2008) found that baby-faced CEOs are perceived as more honest/innocent in times of public relation crises, which helps them shift the blame away from the company more efficiently than mature-faced CEOs. The effect depends on the severity of the crisis, however.

Tanner and Maeng (2012) digitally morphed a stranger's face with the face of a celebrity/politician. Participants judged a stranger's face morphed with Tiger Woods' face as more trustworthy and likable, and a stranger's face morphed with George W. Bush's face as more trustworthy. Importantly, the subjects did not recognize either Woods or Bush in the morphed faces.

With most research on faces in marketing being behavioral, Xiao and Ding (2014) represent the first effort to apply computational analysis to faces in print ads. The authors applied the eigenface method to link facial features of models from print ads from 12 categories to metrics of attitude toward the ad, attitude toward the brand, and purchase intention. The findings show a significant (up to 20%) potential for improving outcome metrics due to face optimization. The face effect was highly heterogeneous by category and target audience, and moderated by the extent to which consumers consider a product hedonic. The proposed data-driven model could be used in practice for "face-screening," that is, selecting faces from the pool of models that match best-performing prototypical features for a specific product and target audience.

The limitations of Xiao and Ding (2014) are as follows. First, they focused on differences between different faces and did not consider cases in which product images have no face.[4] The effect of not having a face in a product image is unknown. Second, they did not focus on models' gazes. Gazes potentially allow more strategic leverage because the same model/face can be used to feature a certain product (and in our shopping context, we indeed found the same model is usually used for a given product). In other words, with facial features kept constant, gaze still can be used to drive the difference in the outcomes. Third, outcome metrics of Xiao and Ding (2014) were behavioral, and examining real-world outcomes from the secondary data, such as browsing, ordering, and returning the products, is important.

We next review existing research on gaze before offering our hypotheses for both face and gaze in our setting.

[4] In the context of our digital shopping data, up to 70% of product images do not have a model's face (§3.4.1).

3.2.2 Existing research on gazes

As a facial region, eyes attract most of the viewer's attention (Sajjacholapunt & Ball, 2014). The ability to encode gazes of others is developed very early in life and serves as an essential building block for social cognition (Frischen et al., 2007). Gaze cues are powerful in steering the viewer's attention (Bayliss et al., 2011). Following the gaze direction of others evolved as an important feature of the human visual system: A person's gaze direction reveals where the person is looking and, subsequently, the future intentions and actions of that person (Clifford & Palmer, 2018).

Gaze cues are also instrumental to emotional expression (Frischen et al., 2007). Subjects were faster in recognizing anger and joy when the model's gaze was direct (i.e., gazing at the viewer) and faster in recognizing fear and sadness when the model's gaze was averted (Adams & Kleck, 2003). Also, the perceived intensity of facial joy/anger was reinforced by direct gaze, whereas the intensity of fear/sadness was reinforced by averted gaze (Adams & Kleck, 2005).

Interestingly, the relationship between gaze direction and emotional expression can be two-way. Specifically, the emotional expression of a model's face can affect how others perceive the gaze direction of the model. When the model's face was happy, respondents were more likely to perceive that model's gaze as directed at them, compared to neutral, fearful, and angry facial expressions (Lobmaier et al., 2008). Martin and Rovira (1982) documented a similar effect for the case of smiling models.

Psychological effects of gaze direction on viewers' perceptions and behavior can potentially be stronger in marketing context. Consider product images in the apparel category. Usually, the same model appears in several images for the same product. Although facial features (and, often, emotional expressions) of the model are fixed, gaze direction can still be freely altered for different images of the same product, hence providing extra strategic leverage.

In marketing, existing studies of gaze-direction effects are predominantly behavioral lab-based. Literature often distinguishes between “direct” gaze (i.e., a model in the product image looks directly at the viewer) and “averted” gaze. A common finding is that the model gazing at the product (as opposed to direct gaze) generates better results for marketers.

Hutton and Nolte (2011) used eye-tracking to measure dwell time on the product and brand regions of a print ad as a function of where the model in the ad is gazing (at the product vs. at the viewer, see Figure 3.1). When the model gazed at the product, the reader’s dwell time was higher for both product and brand regions.

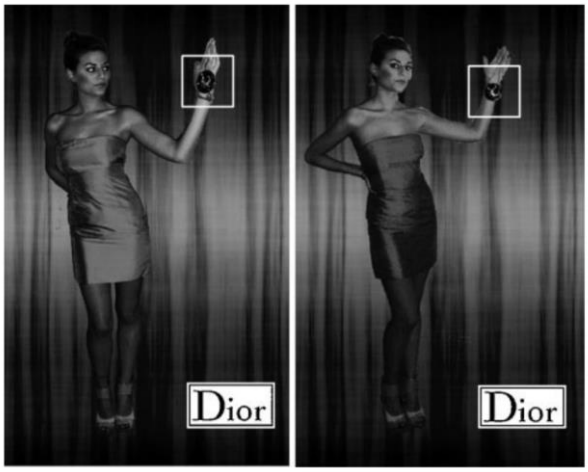


Figure 3.1. Example of Model Gazing at the Product versus at the Viewer.
Note. Adapted from Hutton and Nolte (2011).

Adil et al. (2018) found that a model’s gaze at the product (vs. at the viewer) generates more attention to and memorization of both product and the brand, as also results in better purchase intentions for the product. Palcu et al. (2017) also documented a positive effect of at-product gaze on purchase intentions.

Sajjacholapunt and Ball (2014) studied face and gaze effects in the context of banner ads. They compared three types of ads: (1) without a model’s face, (2) with face and direct gaze, and

(3) with face and gaze directed at the product (averted gaze). Participants spent more time looking at ads with a model's face. When the model's gaze was directed at the viewer, participants dwelled more on the model's face. When the model's gaze was directed at the product, participants dwelled more on the text and product information, and showed better information recall. These results suggest positive effects of at-product gaze on consumer attention and memory.

In the context of social media, Berger and Barasch (2018) examined the potential advantage of candid photos. They distinguished between posed images (direct gaze) and candid images (averted gaze). In a candid photo, a person is not aware of a picture being taken (see Figure 3.2). Candid photos received higher scores on friendship potential, dating potential, overall connectedness, and liking. The authors suggest this "candid advantage" can be explained by genuineness / authenticity of candid photos.

Fig 3.2. Example of a Candid versus Posed Gaze.

Note. Adapted from Berger and Barasch (2018).

Previous studies on gaze effects in marketing have the following limitations:

1. They were small in scale. They used few photos (stimuli) and hence were unable to examine potential moderators (price, category) and heterogeneity. Modern scalable

computer vision and deep learning tools for face and gaze detection and extraction can help overcome this limitation.

2. Previous studies looked at behavioral outcome metrics (dwell time, consumer memory, purchase intentions). No study examined gaze effects on actual business outcomes such as clicks, orders, and returns. For example, consumers can better recall product-related information if the gaze in the banner ad is directed at the product (as opposed to at the viewer, Sajjacholapunt & Ball (2014)), but are the consumers also more likely to click on this ad?
3. Previous research did not distinguish between subtypes of averted (/candid) gaze, such as downwards or sideways. Gaze psychology would suggest, for instance, that "downwards gaze can signal shame or embarrassment" (Clifford & Palmer, 2018, p. 11), which might have new implications for understanding averted-gaze effects.

3.2.3 Hypotheses

Given Sajjacholapunt and Ball's (2014) findings that participants dwelled longer on the ad with a model's face (than on an ad without the face) and Xiao and Ding's (2014) findings that variation in models' faces can drive up to 20% of the difference in the behavioral outcomes, we developed the following hypotheses in our e-commerce setting:

H1. *Product image with a face (vs. without) increases browsing of the product page.*

Next, given findings on underperformance of direct gaze (compared to at-product gaze, Adil et al., 2018, Palcu et al., 2017) on both memorization/attention and the purchase intentions, we expected that

H2. *The model's direct gaze in a product image decreases browsing and ordering of the product.*

Finally, given the notion in psychology that "downwards gaze can signal shame or embarrassment" (Clifford and Palmer 2018, p. 11), we expected that

H3. *The model's downwards gaze in a product image decreases browsing and ordering of the product.*

We next proceed to a description of the data, description of the face- and gaze-extraction model, and the empirical results.

3.3 Description of the data

Our data come from one of the largest Chinese e-commerce websites. We obtained the data on 57,088 apparel products for May through November 2017. Apparel categories include women, men, children apparel, shoes, and accessories.

Each of the products has a webpage with descriptions and images. After submitting a search query on the website, the user scrolls through the results page (Figure 3.3). After clicking to browse the product, the user lands on the product page (Figure 3.4). The product page contains the product title, price, % items claimed, shipping details, size choice, order button, and the product images (median number of images per product is 6).

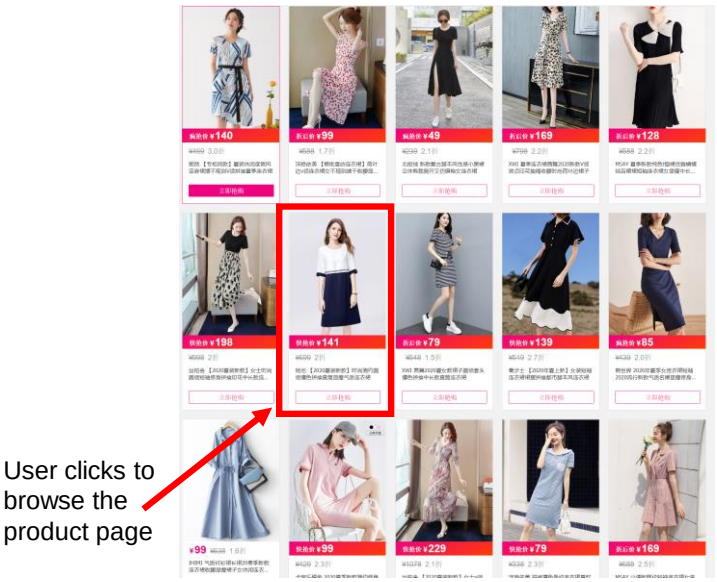


Figure 3.3. Example of Search Results Page e

Fig 3.4. Example of Product Page

Dependent variables included (1) aggregate number of times a product page was browsed, (2) aggregate number of times a product was ordered, and (3) overall amount of product returns. We define two additional dependent variables as follows:

$$Conversion\ rate = \frac{Total\ product\ orders}{Total\ times\ browsed} \tag{3.1}$$

$$Product\ return\ rate = \frac{Total\ product\ returns}{Total\ product\ orders} \tag{3.2}$$

We report summary statistics of the dependent variables in the Table 3.1. The distributions of browses, orders, and returns is highly skewed; hence, we log-transformed these variables for subsequent analysis (see Figure 3.5).

Table 3.1. Summary Statistics of Dependent Variables

VARIABLES	N	Mean	SD	Min	P5	P50	P95	Max
Browses	57,088	524.8	1,305	1	24	183	2,001	47,640
Orders	57,088	10.17	20.00	0	1	4	41	199
Conversion rate	57,088	0.0323	0.0448	0	0.00529	0.0219	0.0870	1
Returns	57,088	1.219	2.986	0	0	0	5	73
Return rate	57,083	0.143	0.246	0	0	0	0.750	1
Log(1+Browses)	57,088	5.285	1.331	0.693	3.219	5.215	7.602	10.77
Log(1+Orders)	57,088	1.761	0.997	0	0.693	1.609	3.738	5.298
Log(1+Returns)	57,088	0.485	0.662	0	0	0	1.792	4.304

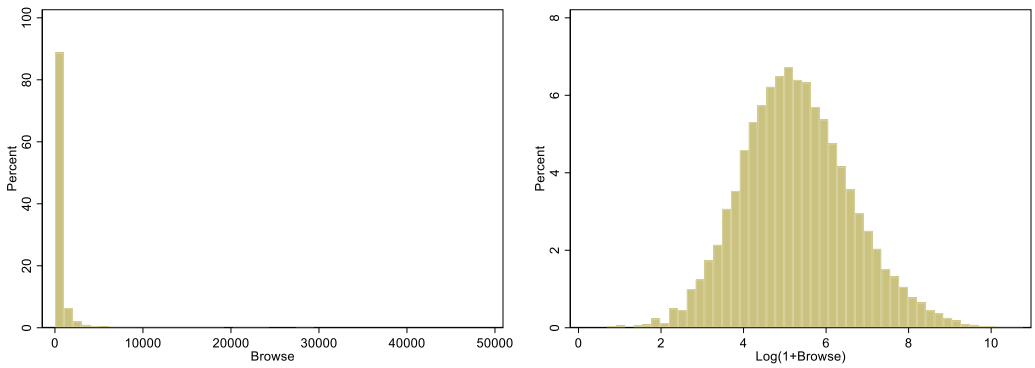


Figure 3.5. Density of "Browse" Variable, Raw (Left) versus Log-Transformed (Right)

We next describe the independent variables of interest pertaining to face and gaze characteristics of the models in the product images.

3.4 Face and gaze extraction: A deep learning approach

A simple pipeline of automated face and gaze detection is as follows. First, a face region (if any) is detected from the image. If a face region cannot be detected, the image is assumed to not have any face present. Second, eye location is identified from the face region. Third, gaze vector is inferred given eye location. We used two separate deep-learning models for face detection and gaze inference, which we describe below.

3.4.1 Face detector

To detect a face in an image, we used the pre-trained convolutional neural net[5] (CNN) that has over 99% accuracy on the "Labeled faces in the wild" dataset (a popular benchmark dataset for face-recognition algorithms, Huang et al. 2007). The CNN takes an image as an input and produces coordinates of the face region as an output (see Figure 3.6). The algorithm performs equally well detecting faces when the model in the image has sunglasses on. We found that accuracy of the CNN-based approach in our setting is much higher than approaches using older and more traditional face detectors, such as Viola-Jones detector (e.g., Lu, Xiao, & Ding, 2016).

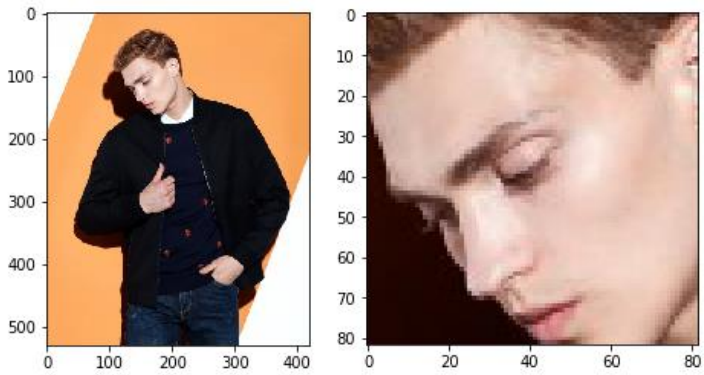


Figure 3.6. Example of Face-Region Detection via CNN in the Shopping Context

Note. The left image represents an input original image; the right image represents face region.The face is located at the pixel location {Top: 72, Left: 119, Bottom: 154, Right: 201}.

[5] https://github.com/ageitgey/face_recognition

Overall, we detected faces in 30% of the photos in our dataset. Face presence varies significantly by category: Although the "shoes" category has a few product photos with faces, almost every product image in "women's/men's top" and "hat" categories include a face (see Table 3.2). Category dummies explain 63% of the variation in the face-presence indicator.

Table 3.2. Summary Statistics of Business Outcomes and Face-Presence Indicator by Category

Category	Frequency	Aggregate browsing (median)	Conversion rate (median)	Return rate (median)	Face present (mean)
Women's top	8%	307	1.1%	20.0%	91%
Men's top	6%	211	2.1%	6.7%	90%
Hat	8%	121	2.4%	0.0%	88%
Scarf	4%	77	3.4%	0.0%	82%
Women's underwear	1%	168	3.5%	0.0%	66%
Eye glasses	3%	112	2.0%	0.0%	61%
Kids' apparel	5%	204	3.5%	0.0%	30%
Unclassified	2%	168	2.2%	0.0%	28%
Women's accessories	12%	199	2.0%	0.0%	15%
Men's pants	2%	168.5	2.5%	8.8%	9%
Handbag	7%	270	1.6%	0.0%	6%
Teapot	1%	81.5	4.5%	0.0%	5%
Women's pants	2%	192	1.6%	22.2%	5%
Umbrella	10%	142	3.2%	0.0%	5%
Other accessories	1%	70	3.5%	0.0%	4%
Watch	2%	126	1.9%	0.0%	2%
Cup	3%	128	4.3%	0.0%	1%
Women's socks	2%	62	6.3%	0.0%	1%
Kids' shoes	2%	228.5	2.0%	0.0%	1%
Women's shoes	13%	449	1.7%	14.3%	1%
Men's shoes	3%	169	1.8%	0.0%	0%
Travel bag	5%	112.5	2.3%	0.0%	0%

Note. Frequency represents percent of the sample (57,088 observations total). Aggregate browsing denotes number of times a product page was browsed. Conversion rate = (total orders) / (total times browsed). Return rate = (total returns) / (total orders). Face-presence indicator extracted using CNN for face detection.

We also computed a facial prominence variable as follows:

$$Facial\_prominence = \frac{Face\ region\ area}{Overall\ picture\ area.} \quad (3.3)$$

We retained a face-presence indicator and facial-prominence variable for the subsequent empirical analysis. If no face is present in the image, the facial-prominence variable equals o.

3.4.2 Gaze extractor

The ability to follow the gazes of others is a fundamental ability of human vision, yet the gaze-direction detection problem has received limited attention in the computer-vision community. Recasens et al. (2015) made an early effort to train a gaze-following model in an unconstrained environment. Previous research only focused on specific contexts such as people looking at each other, directly at the camera, and so on.

The gaze-following model consists of two pathways: the saliency pathway and gaze pathway (Figure 3.7). In the gaze pathway, the model uses the head-only part of the image to create the "gaze mask" (i.e., distribution of likely gaze concentration in the image). In the saliency pathway, the model uses the whole image (excluding the person's location) to generate a saliency map (i.e., the likely distribution of an external viewer's attention over image pixels). The logic is that if a certain object in an image is salient for the external viewer (e.g., a baseball), it is likely to be salient for a person inside the image as well (e.g., baseball player). The gaze mask and the saliency maps are then combined via element-wise product. The product is passed to the final refinement stage ("shifted grids") before making the final prediction of the gaze vector. Figure 3.8 presents the output of each prediction stage.

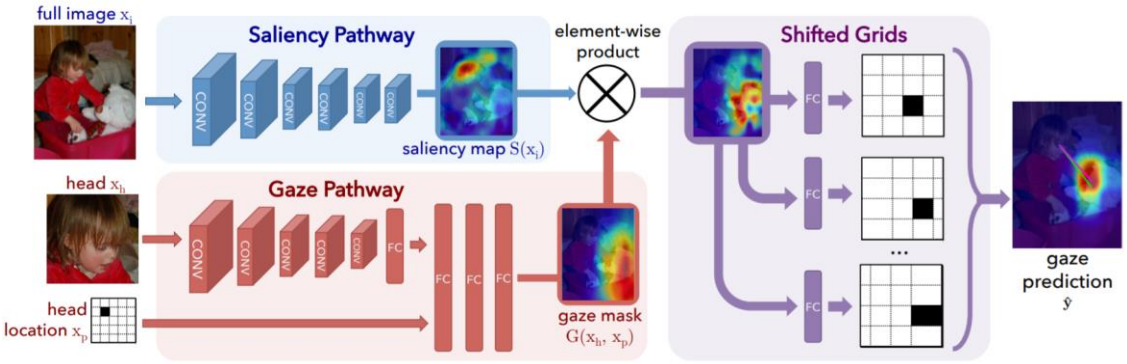


Figure 3.7. Gaze-Following Network Rrchitecture

Note. Adapted from Recasens et al. (2015).

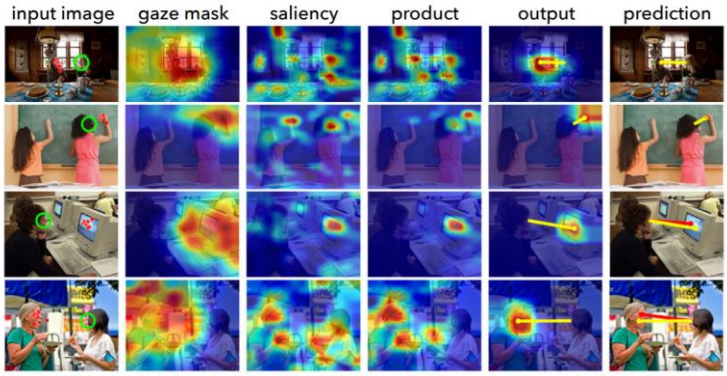


Figure 3.8. Output Visualizations at Each Stage of the Gaze-Following Model

Note. Adapted from Recasens et al. (2015). Green circles denote head position of the focal person. Red dots denote ground-truth annotations of gaze position of the focal person. Red line denotes ground-truth gaze vector. Yellow line denotes predicted gaze vector.

As Figure 3.7 shows, the gaze mask and saliency map complement each other to generate a more precise gaze vector. Lian, Yu, and Gao (2018) noted that Recasens et al. (2015) treated the gaze and saliency pathways as two independent partitions of one problem. Lian, Yu, and Gao (2018) suggested a sequential model that more closely mimics the way humans detect others' gaze directionss. In this model, the gaze pathway would precede the saliency pathway rather than going in parallel. The modified model resulted in better accuracy than Recasens et al. (2015). Because

Lian, Yu, and Gao (2018) also provided the faster GPU implementation, we used their model to detect gaze vectors in the shopping context of interest.

Conceptually, in our empirical study, we distinguished between the four different gaze directions (Figure 3.9): direct, left, right, and downwards. According to the previous literature, the three latter gaze types are labeled "averted gaze." Our study is the first to decompose averted gaze into left, right, and downwards.

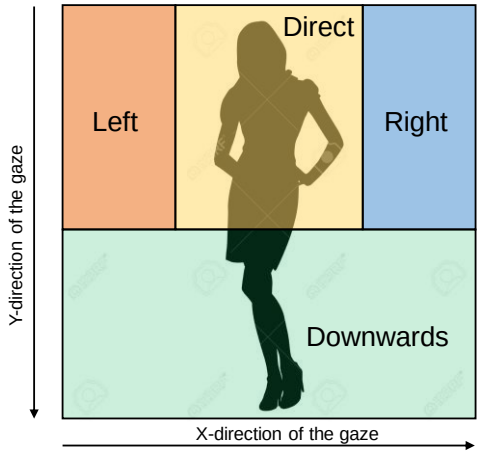


Figure 3.9. The four gaze directions.

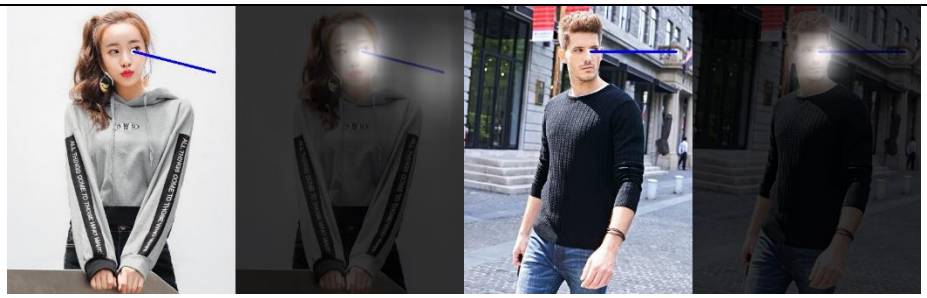
We applied the gaze-direction-detection model (Lian, Yu, & Gao, 2018) to images in our shopping data. Table 3.3 presents example results. The output of the gaze-following model is (X,Y)-coordinates of the gaze location in an image. By connecting original eye location (which is a byproduct of face detection described in section 3.4.1) to the predicted gaze location, we obtained a gaze direction vector. It is highlighted blue in the images in Table 3.3. We observe that "direct" gazes have a very short length of the gaze vector. On the contrary, downwards and sideways gazes are considerably long. Appendix B contains additional examples of detected gazes.

Table 3.3. Examples of the Four Gaze Directions Detected in the Product Images

Panel A. Direct gaze

Panel B. Downwards gaze

Panel C. Left gaze



Panel D. Right gaze

Note. Blue line connects eye-location coordinates to gaze-location coordinates, and represents the gaze-direction vector. Black-and-white counterparts to the images represent gaze-field densities and are an auxiliary output from Lian, Yu, and Gao's (2018) deep-learning model.

Using information about eye coordinates and gaze coordinates, we computed gaze length as a simple Euclidean distance as follows:

$$Gaze_{length} = \sqrt{\left(Eye_{loc_X} - Gaze_{loc_X}\right)^2 + \left(Eye_{loc_Y} - Gaze_{loc_Y}\right)^2}. \quad (3.4)$$

Combining information about gaze length and the (X,Y)-coordinates of the gaze, we computed gaze variables to be used in the empirical exercise. For the downwards gaze, we used the Y-component of the gaze coordinate. For left and right gazes, we used the X-component of the gaze coordinate. Direct gaze is defined as an indicator function of gaze length and the X-coordinate of the gaze as follows:

$$Direct_{gaze} = I\{\left|Gaze_{loc_X} - 0.5\right| < 0.1 \; and \; Gaze_{length} < 0.25\}. \quad (3.5)$$

Hence, the gaze is classified as "Direct" if it is sufficiently short in length and if the X-coordinate of the gaze is sufficiently close to the center of the image. The function in equation 3.5 was tuned based on the visual checks of accuracy of the direct-gaze classification.

As a summary statistic of gaze directions in our dataset, we plot densities of eye locations and gaze locations in Figure 3.10. For eye locations (Figure 3.10, left), we detect a major cluster in the top-center position. This cluster corresponds to predominant location of a model's head in a picture in our data. The second, slightly lower cluster corresponds to the head location in the "Hat" category. For gaze locations (Figure 3.10, right), we detect four clusters corresponding to four gaze directions considered (direct, downwards, left, right).

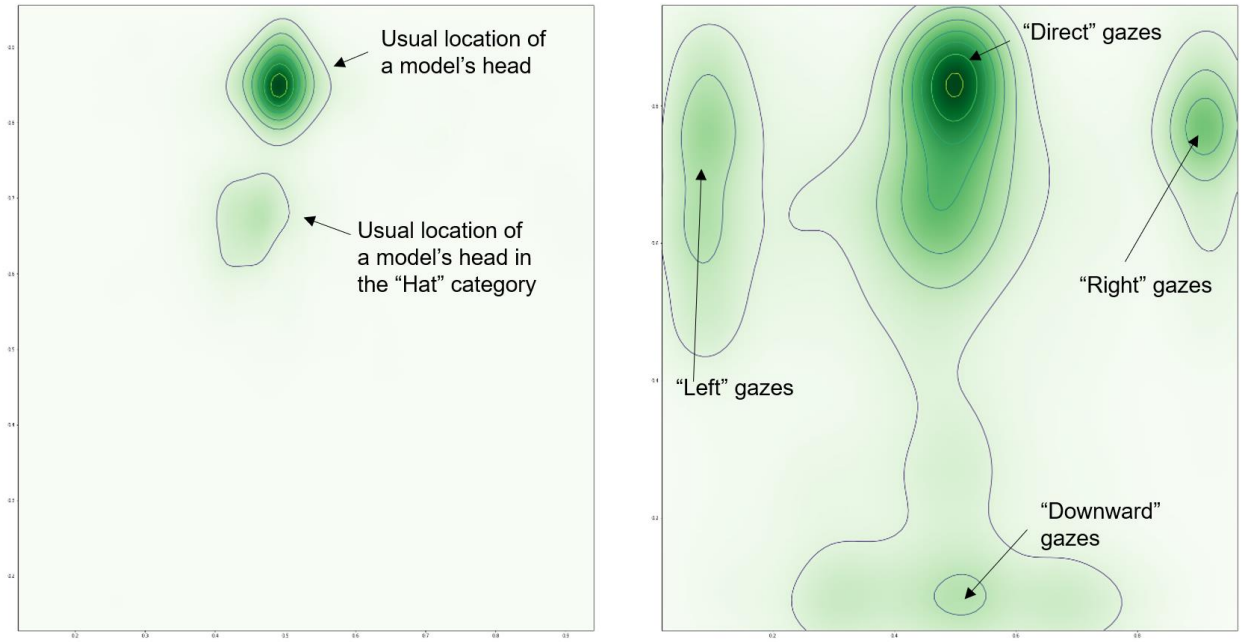


Fig 3.10. Densities of Eye Locations (Left) and Gaze Locations (Right)

Note. Eye locations extracted using face-detection CNN (section 3.4.1). Gaze locations extracted using the gaze-following model by Lian, Yu, and Gao (2018).

As discussed in section 3.2.2, prior literature suggests the potential disadvantage of "direct gazes," yet in our empirical setting, "direct gazes" are used in 30% of the images that have a face. Evaluating the effectiveness of different gaze types in the shopping context is important. In the next section, we link our face and gaze variables to important business metrics.

3.5 Empirical analysis

In this section, we examine the relationship between business metrics of interest (browsing, ordering, conversion rate, and returning) and the face and gaze characteristics of models in the product images, controlling for an extensive set of other product attributes. We employ regression analysis with brand fixed effects at a product level. Our first empirical model for product i follows:

$$Y_i = c + brand_{FE_i} + Price_i + X_i + face_{present_i} + face_{prominence_i} + e_i. \quad (3.6)$$

Here, for product i,

- Y_i denotes the dependent variable of interest. We estimate separate models for (1) Log(1+Browse), (2) Log(1+Order), (3) conversion rate, and (4) Log(1+Return). Description and summary statistics for the variables are in section 3.3.
- c denotes the intercept.
- $brand\_FE_i$ denotes brand fixed effect. The dataset contains 351 unique brands. One brand can serve multiple categories of products.
- $Price_i$ denotes the product price in yuan. Our dataset covers moderately priced products, with the median price being ~20 USD, the average price being ~30 USD, and the maximum price being ~310 USD.
- X_i denotes other product characteristics (described in more detail below).
- $face\_present_i$ denotes the binary indicator of face present as identified by the CNN model (section 3.4.1).
- $face\_prominence_i$ denotes the continuous (0-1) variable defined in equation (3.3).
- e_i is an idiosyncratic error term.

Product characteristics X_i include the following variables:

- Dummy for the date when the product page went online.
- Hour and minute of product-page upload (dummies)
- Colors identified from text product descriptions. Overall, we include indicators for 12 colors (blue, grey, green, yellow, orange, red, pink, purple, brown, black, navy, colorful).
- Product attributes identified from bag-of-words analysis on product text descriptions. Sixty-four indicators are included. Examples of the attributes include fashion, cotton,

comfort, breathable, outdoor, leather, summer, light, and so on.

- Product categories (from Table 3.2). Category dummies are identified in the brand fixed-effects model because a brand can have products in multiple categories.

Table 3.4 provides the results of estimating equation (3.6). The price coefficient is plausibly negative and strongly significant. Face presence has a positive impact on browsing and returns, but an insignificant impact on ordering. As a result, face presence is negatively correlated with the conversion rate. Interestingly, face prominence has a negative impact on browsing but a positive impact on ordering. As a result, face prominence is associated with a higher conversion rate. In terms of goodness of fit, our best model (browsing) explains 48% of the variation, and our worst model (conversion rate) explains 26% of the variation.

Table 3.4. Empirical Results for Face Variables

VARIABLES	(1) Log(Browse)	(2) Log(Order)	(3) Conv rate	(4) Log(Return)
Colors	yes	yes	yes	yes
Categories	yes	yes	yes	yes
Product attributes	yes	yes	yes	yes
Brand FE	yes	yes	yes	yes
Price	-0.00***	-0.00***	-0.00***	-0.00***
	(-18.37)	(-37.22)	(-7.70)	(-12.94)
Face presence	0.12***	-0.03*	-0.00***	0.03**
	(6.01)	(-1.65)	(-5.78)	(2.32)
Face prominence	-0.57***	0.38**	0.04***	0.11
	(-3.01)	(2.40)	(4.68)	(1.13)
Constant	5.23***	1.90***	0.04***	0.48***
	(225.29)	(89.13)	(34.35)	(32.99)
Observations	57,088	57,088	57,088	57,088
R-squared	0.48	0.34	0.26	0.29
Adjusted R-squared	0.476	0.331	0.256	0.283

Note. Robust t-statistics in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Next, we add the gaze variables and estimate the following model:

$$Y_i = c + brand\_FE_i + Price_i + X_i + face\_present_i + face\_prominence_i + Gaze\_loc\_X_i + Gaze\_loc\_Y_i + Direct\_gaze_i + e_i \quad \text{(Eq 3.7)}$$

Here, for product i,

- $Gaze\_loc\_X_i$ denotes the X-coordinate of the gaze location as identified by the CNN model (section section 3.4.2). The variable is continuous (0-1) and a higher value means more "to the right."
- $Gaze\_loc\_Y_i$ denotes the Y-coordinate of the gaze location as identified by the CNN model (section section 3.4.2). The variable is continuous (0-1) and a higher value means "lower/down."
- $Direct\_gaze_i$ – binary indicator of whether gaze is identified as "direct" (i.e., model looking at the viewer) as per function from equation (3.5).
- Other variables defined as before (equation (3.6)).

Table 3.5 presents the results of estimating equation (3.7). Face effects remain largely unchanged compared to Table 3.4. Direct gaze negatively affects both browsing and ordering (impact on conversion rate is not significant). This result is consistent with previous behavioral studies examining direct- versus averted-gaze effects (section 3.2.2). Next, gazes that are directed more to the right (vs. left) are associated with higher browsing. Effect on orders is insignificant. Finally, downward-directed gazes negatively affectt both browsing and ordering (impact on conversion rate is not significant). Both the direct gaze and downwards gaze are associated with fewer product returns.

Table 3.5. Empirical Results for Face and Gaze Variables

VARIABLES	(1) Log(Browse)	(2) Log(Order)	(3) Conv rate	(4) Log(Return)
Colors	yes	yes	yes	yes
Categories	yes	yes	yes	yes
Product attributes	yes	yes	yes	yes
Brand FE	yes	yes	yes	yes
Price	-0.00***	-0.00***	-0.00***	-0.00***
	(-18.46)	(-37.23)	(-7.65)	(-12.99)
Face presence	0.15***	0.04	-0.00*	0.05***
	(4.91)	(1.61)	(-1.86)	(2.63)
Face prominence	-0.53***	0.39**	0.04***	0.13
	(-2.80)	(2.51)	(4.61)	(1.24)
Direct gaze indicator	-0.10***	-0.08***	0.00	-0.03**
	(-5.41)	(-4.84)	(1.34)	(-2.26)
X-direction of the gaze (higher means right)	0.13***	0.01	-0.01***	0.02
	(3.97)	(0.26)	(-4.65)	(1.06)
Y-direction of the gaze (higher means down)	-0.16***	-0.10***	0.00	-0.06***
	(-4.81)	(-4.02)	(0.24)	(-2.91)
Constant	5.23***	1.90***	0.04***	0.48***
	(225.39)	(89.11)	(34.34)	(33.00)
Observations	57,088	57,088	57,088	57,088
R-squared	0.48	0.34	0.26	0.29
Adjusted R-squared	0.476	0.331	0.256	0.284

Note. Robust t-statistics in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

The decomposition of the averted gaze into downwards and sideways gazes complements previous research on averted-gaze effects. Previous studies found that averted, product-directed gaze increases viewers' dwell time on the product (e.g., Adil et al., 2018; Palcu et al., 2017). We found that, in the apparel context, the downwards gaze is suboptimal even if it is product directed. A potential explanation could be that downwards gaze might be associated with shame and embarrassment (Clifford & Palmer, 2018), hence resulting in fewer clicks and orders. The examination of the exact mechanism of downwards-gaze effects is a direction for future research.

The results suggest the following implications for product photo designers. Overall, featuring a model's face in a product image is beneficial. Face prominence should be higher to achieve more orders and a better conversion rate, but lower to achieve more product browses. Next, a model's gaze should be neither direct nor downwards. A sideways gaze is a solution, and a right-directed gaze works somewhat better than a left-directed gaze.

3.6 Discussion

This study is the first attempt to quantify both face and gaze characteristics of the models in product images, and link them to real business outcomes (browses, orders, product returns) using a large-scale secondary dataset on online shopping. We replicated previous behavioral results that a "direct" gaze generally leads to worse results than an "averted" gaze. However, we distinguished between subtypes of "averted" gazes (downwards vs. sideways) and demonstrated that the downward gaze generates worse results than the sideways (particularly, to-the-right) gaze. Extant empirical and behavior literature do not document the results for downwards and sideways gazes. We suggest a potential explanation of underperformance of the downwards gaze stemming from potential associations with shame and embarrassment (Clifford & Palmer, 2018), but the investigation of the exact mechanism is an avenue for future research. We also report actionable insights for product image design. Further directions for research include examining the heterogeneity of face and gaze impacts by category, and examining the role of potential moderators (e.g., price).

Concluding remarks and future research directions

Despite the rise of the visual web, little academic research has been devoted to study the impact of visual strategies on engagement with firm-generated visual content and to assess the relative effectiveness of text-based versus image-based content strategies. This dissertation addresses this gap, suggesting a framework (and developing tool) to measure and compare the text and images of a brand's social media messages on the dimensions of sentiment and arousal (activation/motivation to act), the two most prominent dimensions in affective science. This dissertation is also the first attempt to quantify both face and gaze characteristics of the models in product images, and link them to real business outcomes (browses, orders, product returns) using a large-scale secondary dataset on online shopping.

In essay 1, we proposed and validated a tool to predict an emotion of a digital image given purely pixel-level information from the image. We extractdc visual characteristics from the images that correspond to four emotion modalities: (1) elements of design (low-level visual features) such as color, texture, shape, lines, curves, corners, edges, and orientation; (2) high-level visual objects/concepts (e.g., adventure, action, leisure, danger, etc.); (3) human facial expressions; and (4) text embedded in the image. We linked the four modalities to the emotional dimensions of the image, using the gradient boosting algorithm. The accuracy is over 80%, and the model has near-perfect predictions for images with extreme ground-truth levels of Sentiment/Arousal. The demo of the visual-emotion prediction tool is available at imagesentiment.com.

Next, we reported and validated individual elements of design - drivers of visual emotion. For example, from the predictive model, we found that color variety, non-smooth texture, higher amounts of green and orange hues in an image, higher brightness, and horizontal orientation are associated with more positive sentiment of the image. Many corners, red, and pink hues are

associated with high arousal, whereas blue hues are associated with low arousal. We validated each of the independent drivers in the lab by modifying original firm-generated tweet images to have less (or more) of a particular visual feature. For 77% of the image pairs, we found statistically significant differences in the scores assigned by human subjects, and these differences agree with model predictions.

Further directions for essay 1 include examining the heterogeneity in the viewer's perceptions of visual emotions. Linking the patterns of how subjects annotate visual sentiment/arousal to these subjects' characteristics would be of interest. Understanding how the target audience perceives the emotionality of visual content is important for a social media marketer. Second, we can examine heterogeneity in individual drivers of visual emotion. We found that, for example, on average, color variety leads to more positive sentiment, but we did not examine the potential heterogeneity. This might be particularly relevant for visual features that have different cultural associations (e.g., red). Third, the interactions between emotion modalities (e.g., red+child) could be examined in more detail in their impact on visual emotions.

In essay 2, we undertook an empirical analysis of engagement (Retweeting) with 1.3 million tweets from 630+ brands from 11 categories, posted since 2008. We used a visual-emotion scoring tool to create variables of visual sentiment and arousal for Tweet images. For Tweet-text sentiment, we used an existing tool (Valence Aware Dictionary for sEntiment Reasoning (VADER)), optimized specifically for Twitter posts. Tweet text is higher on arousal/activation score if it uses exclamation marks, question marks, a high proportion of uppercase letters, and/or "call-to-action" verbs (e.g., try, share, buy, etc.). We examined the impact of image/text strategies on Retweeting using an"brand — week-of-the-year" fixed-effects regression with multiple controls such as Tweet-text meaning, high-level visual objects, text embedded in images, logo color

distance, and so on. The full model explains 70% of the variation in Retweeting. We found the effectiveness of the high-arousal/motivation-to-act text strategy has decreased over time, with the impact switching from positive to negative sometime around 2011. We suggest this finding could be explained by Friestad and Wright's (1994) persuasion knowledge model: Over time, consumers have become significantly more resistant to high-activation text. Interestingly, we did not find the same pattern for high-activation imagery, with the high-aarousal image strategy staying dominant for images that also exhibit negative visual sentiment. These results suggest *marketers should* ***decrease activation through text and instead leverage imagery to activate consumers*** *in their social media communications*. Further, we found significant differences in dominant visual strategies by sector, with, for example, "negative-sentiment" images being dominant for charities/non-profits, and "positive-high-arousal" images being dominant for quick-service restaurants. We estimate the industry-specific dominant image strategy has the potential to increase engagement by up to 19%, whereas the text strategy has the potential to increase engagement by up to 14%.

Further efforts should focus on a more granular examination of the heterogeneity of the text/image engagement impacts. We did not zoom in on the categories to understand brand-specific heterogeneity. Given that a single brand may upload thousands of Tweets, the data should be rich enough to examine brand-level heterogeneity. Another direction for research is further investigation of the psychological mechanism behind the effects. We suggest the interpretation of the effects based on persuasion knowledge and elaboration likelihood models, but our findings do not causally prove this mechanism. Finally, we considered Retweets as an engagement metric, but it looking at alternative metrics, such as likes, comments, follower base, might be interesting.

Essay 3 is the first attempt to quantify both face and gaze characteristics of the models in product images, and link them to real business outcomes (browses, orders, product returns) using a large-scale secondary dataset on online shopping. We utilized deep-learning algorithms for face detection and gaze-following in the context of 57,088 apparel products from 22 categories from one of the largest Chinese e-commerce websites. We extracted characteristics such as face presence and prominence in product images, and classified the model's gaze into "direct," "downwards," and "sideways." We found that product images that include the model's face are browsed more. However, higher prominence of the face leads to less browsing, but more ordering of the product. Next, we documented a negative impact of the "direct" gaze (i.e., the gaze of the model directed toward the viewer) on both browsing and ordering. This result is consistent with previous behavioral findings but was never demonstrated in an empirical setting. Next, we distinguished between subtypes of "averted" gazes (downwards vs. sideways) and demonstrated that the downward gaze generates worse results than the sideways (particularly, to-the-right) gaze. Extant empirical and behavioral literature do not document the results for downwards and sideways gazes. We suggest a potential explanation of underperformance of the downwards gaze stemming from potential associations with shame and embarrassment (Clifford & Palmer, 2018).

Future research can focus on the investigation of the exact mechanism of the underperformance of the downwards gaze. Other directions for research include examining the heterogeneity of face and gaze impacts by category, and examining the role of potential moderators (e.g., price).

Overall, the computer-vision-based marketing research is an exciting, dynamic, and promising subfield. The visual content (images, videos) offers numerous potential variables to be extracted and linked to important outcomes such as consumer engagement, demand, and so on.

This research is also highly relevant for the practitioner because it offers (1) tools on how to analyze visual content and (2) simple insights on how to use the visuals. The applications of the ideas from this dissertation to the video (rather than static imagery) context (e.g., time-varying sentiment in videos, or dynamic model gazing in product videos) might generate new non-trivial insights for social media marketers and product photographers.

References

Adams Jr, R. B., & Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological science*, 14(6), 644-647.

Adams Jr, R. B., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1), 3.

Adil, S., Lacoste-Badie, S., & Droulers, O. (2018). Face Presence and Gaze Direction In Print Advertisements: How They Influence Consumer Responses—An Eye-Tracking Study. *Journal of Advertising Research*, 58(4), 443-455.

Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of consumer research*, 34(4), 468-479.

Akpinar, Ezgi, and Jonah Berger (2017), "Valuable virality," *Journal of Marketing Research* 54, no. 2: 318-330.

Arnheim, R. (1965). Art and visual perception: A psychology of the creative eye. University of California Press.

Barbosa, L., & Feng, J. (2010, August). Robust Sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). *Association for Computational Linguistics*.

Bayliss, A. P., Bartlett, J., Naughtin, C. K., & Kritikos, A. (2011). A direct link between gaze perception and social attention. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 634.

Belloni, A., & Chernozhukov, V. (2013), "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19(2), 521-547.

Berger, J., & Barasch, A. (2018). A candid advantage? The social benefits of candid photos. *Social Psychological and Personality Science*, 9(8), 1010-1016.

Berger, Jonah, and Katherine L. Milkman (2012), "What makes online content viral?" *Journal of marketing research* 49, no. 2: 192-205.

Borah, A., & Tellis, G. J. (2016) "Halo (spillover) effects in social media: do product recalls of one brand hurt or help rival brands?" *Journal of Marketing Research*, 53(2), 143-160.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (pp. 1-45). Technical report C-1, the center for research in psychophysiology, University of Florida.

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992), "Remembering pictures: pleasure and Arousal in memory," *Journal of experimental psychology*: Learning, Memory, and Cognition, 18(2), 379.

Buter, B., Dijkshoorn, N., Modolo, D., Nguyen, Q., van Noort, S., van de Poel, B., ... & Salah, A. (2011). Explorative visualization and analysis of a social network for arts: the case of deviantART. *Journal of Convergence*, Volume, 2(1).

Campbell, Margaret C., and Amna Kirmani (2000), "Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent." *Journal of consumer research* 27, no. 1: 69-83.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3), 345-354.

Clifford, C. W., & Palmer, C. J. (2018). Adaptation to the direction of others' gaze: a review. *Frontiers in psychology*, 9, 2165.

Corona, H., & O'Mahony, M. P. (2015). A Mood-based Genre Classification of Television Content. arXiv preprint arXiv:1508.01571.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In European Conference on Computer Vision (pp. 288-301). Springer Berlin Heidelberg.

Dondis, D. A. (1974). A primer of visual literacy. MIT Press.

Duffy, E. (1934). Emotion: an example of the need for reorientation in psychology. Psychological Review, 41(2), 184.

Dzyabura, D., El Kihal, S., Hauser, J., & Ibragimov, M. (2018). Leveraging the Power of Images in Managing Product Return Rates. Available at SSRN.

Friestad, Marian, and Peter Wright (1994), "The persuasion knowledge model: How people cope with persuasion attempts," Journal of consumer research 21, no. 1:1-31.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4), 694.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller (2007), "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Technical Report 07-49

Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015), "Multimedia Lab $@ $ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations," in *Proceedings of the Workshop on Noisy User-generated Text* (pp. 146-153).

Gujral, Ranvir (2015), "State of the Industry Visual Marketing: Scale to Win," industry report, https://digiday.com/wp-content/uploads/2015/04/Chute_Digiday_SOTI.pdf

Hashimoto, A., & Clayton, M. (2009). Visual design fundamentals: a digital approach. Nelson Education.

Ho, Jenny (2017), "Machine Learning for Causal Inference: An Application to Air Quality Impacts on House Prices," working paper

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for Sentiment analysis of social media text. In Eighth International AAAI Conference on Weblogs and Social Media.

Hutton, S. B., & Nolte, S. (2011). The effect of gaze cues on attention to print advertisements. *Applied Cognitive Psychology*, 25(6), 887-892.

Jalali, Nima Y., and Purushottam Papatla (2016), "The palette that stands out: Color compositions of online curated visual UGC that attracts higher consumer interaction," *Quantitative Marketing and Economics* 14, no. 4: 353-384.

Khosla, A., Das Sarma, A., & Hamid, R. (2014). What makes an image popular?. In Proceedings of the 23rd international conference on World wide web (pp. 867-876). ACM.

Koshy (2016), "12 Significant Visual Content Marketing Statistics You Need To Know," http://www.sproutworth.com/visual-content-marketing-statistics/?hvid=5bbDNz

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. K. (2016), "From social to sale: The effects of firm-generated content in social media on customer behavior," *Journal of Marketing*, 80(1), 7-25.

Lee, D., Hosanagar, K., & Nair, H. S. (2018), "Advertising content and consumer engagement on social media: Evidence from Facebook," *Management Science*, 64(11), 5105-5131.

Lian, D., Yu, Z., & Gao, S. (2018). Believe It or Not, We Know What You Are Looking At!. In Asian Conference on Computer Vision (pp. 35-50). Springer, Cham.

Liechty, J., Pieters, R., & Wedel, M. (2003). Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika*, 68(4), 519-541.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020), "Visual listening in: Extracting brand image portrayed on social media," working paper

Lobmaier, J. S., Tiddeman, B. P., & Perrett, D. I. (2008). Emotional expression modulates perceived gaze direction. *Emotion*, 8(4), 573.

Lovett, M., Peres, R., & Shachar, R. (2014). A data set of brands and their characteristics. *Marketing Science*, 33(4), 609-617.

Lu, S., Xiao, L., & Ding, M. (2016). A video-based automated recommender (VAR) system for garments. *Marketing Science*, 35(3), 484-510.

Machajdik, J., & Hanbury, A. (2010, October). Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia (pp. 83-92). ACM.

Maeng, A., & Aggarwal, P. (2018). Facing dominance: anthropomorphism and the effect of product face ratio on consumer preference. *Journal of Consumer Research*, 44(5), 1104-1122.

Martin, W., & Rovira, M. (1982). Response biases in eye-gaze perception. *The Journal of psychology*, 110(2), 203-209.

Mayzlin, D., & Yoganarasimhan, H. (2012). Link to success: How blogs build an audience by promoting rivals. *Management Science*, 58(9), 1651-1668.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miniard, P. W., Bhatla, S., Lord, K. R., Dickson, P. R., & Unnava, H. R. (1991), "Picture-based persuasion processes and the moderating role of involvement," *Journal of consumer research*, 18(1), 92-107.

Mitchell, A. A. (1986), "The effect of verbal and visual components of advertisements on brand attitudes and attitude toward the advertisement," *Journal of consumer research*, 13(1), 12-24.

Munsell, Albert H. (1912). "A Pigment Color System and Notation". The American Journal of Psychology. 23 (2): 236–244.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521-543.

Palcu, J., Sudkamp, J., & Florack, A. (2017). Judgments at gaze value: gaze cuing in banner advertisements, its effect on attention allocation product judgments. *Frontiers in psychology*, 8, 881.

Petty, Richard E., and John T. Cacioppo (1986), "The elaboration likelihood model of persuasion," *Advances in experimental social psychology* 19: 123-205.

Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing*, 68(2), 36-50.

Pieters, R., & Wedel, M. (2007). Goal control of attention to advertising: The Yarbus implication. *Journal of consumer research*, 34(2), 224-233.

Pieters, R., & Wedel, M. (2018). Heads up: Head movements during ad exposure respond to consumer goals and predict brand memory. *Journal of Business Research*.

Pieters, R., Warlop, L., & Wedel, M. (2002). Breaking through the clutter: Benefits of advertisement originality and familiarity for brand attention and memory. *Management science*, 48(6), 765-781.

Pieters, R., Wedel, M., & Zhang, J. (2007). Optimal feature advertising design under competitive clutter. *Management Science*, 53(11), 1815-1828.

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of experimental psychology: Applied*, 7(3), 219.

Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking?. In Advances in Neural Information Processing Systems (pp. 199-207).

Rubin, D. C., & Talarico, J. M. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8), 802-808.

Russel, J. A. (1980). A circumplex model of affect. *J. Personality and Social Psychology*, 39, 1161-78.

Sajjacholapunt, P., & Ball, L. J. (2014). The influence of banner advertisements on attention and memory: human faces with averted gaze can enhance advertising effectiveness. *Frontiers in Psychology*, 5, 166.

Stelzner, M. (2016), How Marketers Are Using Social Media to Grow Their Businesses. Social media Marketing Industry report. Social Media Examiner.

Stephen, A. T., Sciandra, M. R., & Inman, J. J. (2015). The effects of content characteristics on consumer engagement with branded social media content on Facebook. *Marketing Science Institute Working Paper Series*, (15-110).

Tanner, R. J., & Maeng, A. (2012). A tiger and a president: Imperceptible celebrity facial cues influence trust and preference. *Journal of Consumer Research*, 39(4), 769-783.

Timoshenko, Artem, and John R. Hauser (2019), "Identifying customer needs from user-generated content." *Marketing Science* 38, no. 1: 1-20.

Tirunillai, S., & Tellis, G. J. (2014), "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation", *Journal of Marketing Research*, 51(4), 463-479.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626.

Tucker, Catherine E. (2014), "The reach and persuasiveness of viral video ads." *Marketing Science* 34, no. 2: 281-296.

Villarroel Ordenes, F., Grewal, D., Ludwig, S., Ruyter, K. D., Mahr, D., & Wetzels, M. (2018), "Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages," *Journal of Consumer Research*, 45(5), 988-1012.

Wang, Z., Mao, H., Li, Y. J., & Liu, F. (2017). Smile big or not? Effects of smile intensity on perceptions of warmth and competence. *Journal of Consumer Research*, 43(5), 787-805.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4), 1191-1207.

Watson, D. and A. Tellegen (1985), "Toward a consensual structure of mood", *Psychological bulletin*, 98(2), p.219.

Wedel, M., & Pieters, R. (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing science*, 19(4), 297-312.

Wedel, M., & Pieters, R. (2017). A review of eye-tracking research in marketing. In Review of marketing research (pp. 123-147). Routledge.

Wedel, M., Pieters, R., & Liechty, J. (2008). Attention switching during scene perception: How goals influence the time course of eye movements across advertisements. *Journal of Experimental Psychology: Applied*, 14(2), 129.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7), 592-598.

Xiao, L., & Ding, M. (2014). Just the faces: Exploring the effects of facial features in print advertising. *Marketing Science*, 33(3), 338-352.

Xiao, Li, and Min Ding (2014), "Just the faces: Exploring the effects of facial features in print advertising." *Marketing Science* 33, no. 3: 338-352.

Yarbus, Alfred L. 1967. Eye Movements and Vision. New York: Plenum Press.

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science*, 66(3), 1045-1070.

Zebrowitz, L. A. (2006). Finally, faces find favor. *Social Cognition*, 24(5), 657-701.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and human behavior*, 15(6), 603-623.

Zhang, J., Wedel, M., & Pieters, R. (2009). Sales effects of attention to feature advertisements: a Bayesian mediation analysis. *Journal of Marketing Research*, 46(5), 669-681.

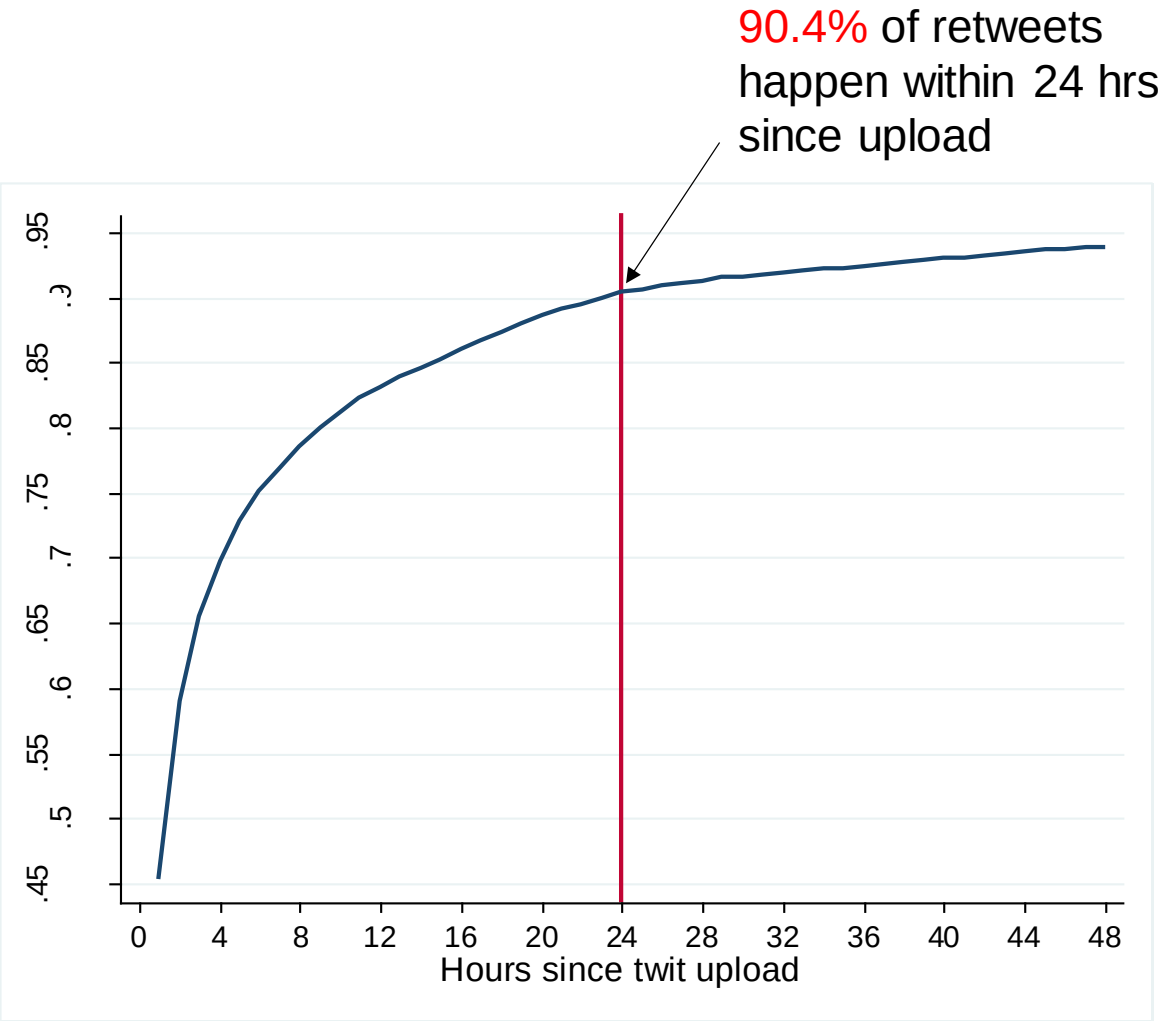
Zhang, Mengxia and Luo, Lan (2018), Can User-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp. Available at SSRN.

Zhang, Shunyuan, Dokyun Lee, Param Vir Singh, Kannan Srinivasan (2017), "How Much Is an Image Worth? Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics", working paper

Appendix A

Using the cumulative number of Retweets as the dependent variable

To examine how quickly the cumulative number of Retweets reaches the "ceiling" for a given Tweet, we ran a side data-scraping exercise. From January 24 to February 20, 2017, we collected Twitter timelines of 80 restaurant brands every hour, giving us the panel dataset at the Tweet-hour level. Overall, we captured 2,741 Tweets that were newly uploaded during the period. Each Tweet was tracked for 340 hours, on average. When examining the cumulative number of Retweets by time, for 90.4% of the Tweets in our sample, Retweets did not increase past the 24 hours since the upload time. We conclude that for the vast majority of tweets, engagement flattens out within 24 hours.



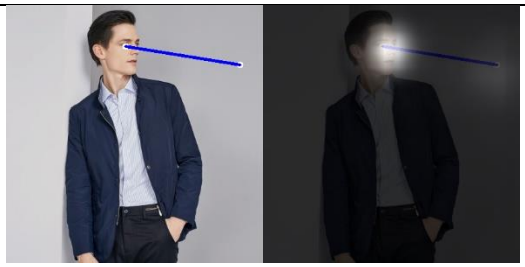
Appendix B

Additional examples of the four gaze directions detected in the shopping data

Panel A. Direct gaze

Panel B. Downwards gaze

Panel C. Left gaze



Panel D. Right gaze

Vita

Evgeny (Eugene) Pavlov was born in Norilsk, Russia. He earned an MA in economics from the Higher School of Economics and New Economic School (Moscow). During 2014-2020, he did a PhD in Marketing at the Univesity of Washington – Seattle.