MIS Quarterly

METHODS ARTICLE

ENHANCING SOCIAL MEDIA ANALYSIS WITH VISUAL DATA ANALYTICS: A DEEP LEARNING APPROACH[1]

Donghyuk Shin
W. P. Carey School of Business, Arizona State University, Tempe, AZ 85281 U.S.A. {dhs@asu.edu}

Shu He
School of Business, University of Connecticut, 2100 Hillside Road, Storrs, CT 06269 U.S.A. {shu.he@uconn.edu}

Gene Moo Lee
Sauder School of Business, The University of British Columbia, 2053 Main Mall,
Vancouver, BC V6T 1Z2 CANADA {gene.lee@sauder.ubc.ca}

Andrew B. Whinston
McCombs School of Business, The University of Texas at Austin, 2110 Speedway,
Austin, TX 78705 U.S.A. {abw@uts.cc.utexas.edu}

Suleyman Cetintas
Advertising Science, Yahoo! Research, 701 First Avenue,
Sunnyvale CA 94089 U.S.A. {cetintas@verizonmedia.com}

Kuang-Chih Lee
Marketplace Optimization, Alibaba Group, 400 S. El Camino Real,
San Mateo, CA 94402 U.S.A. {leekc307@gmail.com}

This research methods article proposes a visual data analytics framework to enhance social media research using deep learning models. Drawing on the literature of information systems and marketing, complemented with data-driven methods, we propose a number of visual and textual content features including complexity, similarity, and consistency measures that can play important roles in the persuasiveness of social media content. We then employ state-of-the-art machine learning approaches such as deep learning and text mining to operationalize these new content features in a scalable and systematic manner. For the newly developed features, we validate them against human coders on Amazon Mechanical Turk. Furthermore, we conduct two case studies with a large social media dataset from Tumblr to show the effectiveness of the proposed content features. The first case study demonstrates that both theoretically motivated and data-driven features significantly improve the model's power to predict the popularity of a post, and the second one highlights the relationships between content features and consumer evaluations of the corresponding posts. The proposed research framework illustrates how deep learning methods can enhance the analysis of unstructured visual and textual data for social media research.

Keywords: Social media, visual data analytics, prediction, machine learning, deep learning, word embedding, image-text similarity

[1]Hsinchun Chen was the accepting senior editor for this paper. Michael Chau served as the associate editor.

Introduction

Social media platforms have attracted billions of users and have emerged as one of the most important channels for companies to communicate with existing and potential customers.[2] Companies have substantially increased their investments and activities in social media marketing,[3] in which visual content plays an important role. The old saying "a picture is worth a thousand words" has never been more true in the era of social media. Images increase the odds of a post getting noticed, especially when people are overwhelmed by the unprecedented amounts of information produced everyday.[4] Industry reports have found concurring results showing that social media posts with images tend to get more likes and shares.[5] Thus a pressing issue for companies conducting social media marketing is to systematically understand the role of visual content in improving customer engagement.[6]

Although social media analysis research has extensively studied textual content (Lakkaraju and Ajmera 2011; Lee et al. 2018; Ma et al. 2015; Stieglitz and Dang-Xuan 2013; Singh et al. 2014) or other characteristics (Lakkaraju and Ajmera 2011; Zadeh and Sharda 2014), visual content has not been systematically investigated due to methodological challenges. In the information systems (IS) or marketing literature, the generation of visual features required significant domain knowledge and manual coding, which is impractical in processing large datasets such as those in social media. In the context of traditional media, studies have shown the impact of images on the effectiveness of ads (Edell and Staelin 1983; Kim and Lennon 2008; Mitchell 1986; Pieters et al. 2010; Pieters et al. 2007; Smith 1991; Tuch et al. 2009; Wu et al. 2016). However, due to methodological restrictions, these studies were based on a small number of hand-picked pictures. Therefore, the findings of these papers can lack generalizability and face scalability issues.

Recently, deep learning emerged as a dominant approach to image recognition in the computer vision field. For example, in the ImageNet image recognition challenge (Russakovsky et al. 2015), performance has dramatically improved due to the breakthrough of neural network-based deep learning approaches. It is believed that, in certain computer vision tasks, deep learning models even surpass human abilities (He et al. 2015; Lake et al. 2015; Rajpurkar et al. 2018; Yu et al. 2016). With such technological advances, we can now generate in a robust and scalable manner visual content features that were previously handcrafted from a small number of images. To the best of our knowledge, a deep learning-based approach has not been widely used to extract visual data features in IS research.

In this research methods article, we introduce deep learning to the social media analysis literature by providing detailed guidance and showing its effectiveness with two empirical case studies. Specifically, we utilize the convolutional neural network (CNN) (Krizhevsky et al. 2017; LeCun et al. 2015), one of the most successful deep learning approaches for visual data analysis. The CNN model aims to automatically discover intricate structure in high-dimensional image data to understand images' semantic content. Our paper provides comprehensive directions on how to conduct a visual content study using deep learning models by summarizing open-source deep learning libraries and outlining steps to construct visual features from images. Additionally, we provide a framework that researchers can use to validate deep learning-induced measures against human coders from Amazon Mechanical Turk (AMT) to ensure their correctness.

In particular, we have generated and validated a group of visual content measures that have important implications in social media research. We first draw on theory from IS and marketing to motivate visual and textual features that can affect consumer information processing, which we operationalize using machine learning (ML) approaches. These features include (1) an object-level image complexity measure based on object detection, (2) an aesthetic score that measures perceived image quality, (3) an adult content score (e.g., underwear, lingerie, etc.), and (4) celebrity endorsements (e.g., Barack Obama, Taylor Swift). In addition, we extract data-driven *generic representations* of visual and textual content that cannot be manually coded (Bengio et al. 2013; LeCun et al. 2015). We then propose novel content features capturing semantic relations between different posts and content types, namely, measures of *content consistency* (e.g., how similar a post is compared to its blog's the average content) and *image-text similarity* (e.g., how closely the image is related to the text within a post). Such measures are difficult to construct without the aid of deep learning approaches, especially for image-text similarity, since pixels and characters are intrinsically different data types.

[2]Statista, "Number of Global Social Media Users" (https://goo.gl/Vxe5uy).

[3]Statista, "Social Media Marketing Spending in the United States from 2014 to 2019" (https://goo.gl/jQjLeb).

[4] In 2017, Facebook and Instagram users posted 195 million and 95 million new photos daily, respectively (see https://goo.gl/VJY51F).

[5]Adobe Social Intelligence Report, Q1 2014.

[6]HubSpot, "42 Visual Content Marketing Statistics" (https://goo.gl/aRJXs3).

To demonstrate the effectiveness of the proposed methods, we conduct two case studies using a large real-life social media dataset. In the first case study, we build ML models using visual and textual content features to *predict* a social media post's popularity. The results show that both theoretically motivated visual features and deep learning-enabled generic representations of images can significantly improve prediction accuracy. In the second study, we empirically investigate how the visual content of a social media post *influences* customer engagement. Drawing on the elaboration likelihood model (ELM) of persuasion (Petty and Cacioppo 1986; Petty et al. 1983), our empirical analyses show that posts including positive peripheral cues and requiring less mental elaboration receive better consumer engagement.

Following best practices of theory-informed big data research approaches (Johnson et al. 2019; Maass et al. 2018; Rai 2016), we summarize our proposed framework for social media research leveraging deep learning with unstructured visual and textual data in Figure 1. Theoretical foundations developed from prior literature inform both feature construction and analysis aimed to address the research question of interest. After features have been identified, generating them from unstructured data sources can be achieved with deep learning methods in a robust and scalable manner. This usually involves an iterative ML model development process, where additional data collection may be necessary depending on the target feature and ML model (e.g., new labels or more granular data). Newly constructed features must be validated through rigorous evaluations in order to be used in the main analysis. Each block in Figure 1 is discussed in the corresponding section.

Our paper makes significant contributions to both academic research and industry practice. Studies on visual content have had limitations due to scalability (i.e., hand-crafted visual features) and generalizability issues (i.e., relatively small-scale studies in controlled environments) (Childers and Houston 1984; Li et al. 2016; Petty et al. 1983; Pieters et al. 2010; Unnava and Burnkrant 1991). For the same reasons, few social media studies have considered the influence of visual content. We aim to fill this research gap by proposing a visual data analytics framework. We demonstrate the process, from variable definitions motivated by theory and driven by big data to feature construction via deep learning models and validation through AMT to application in social media data studies. Our framework can motivate IS researchers to incorporate visual content in empirical studies. From a practical perspective, deep learning models allow companies to generate generic image features in a scalable manner and to gain a better understanding of the impact of images on key social media measures. In other words, our proposed method helps social media managers make informed decisions in ad content engineering.

The remainder of this paper is organized as follows: First, we build a theoretical foundation to motivate visual and textual content features. We then briefly introduce the concepts and background of deep learning models, and construct visual and textual content features using ML models. We validate the newly constructed measures using AMT. Based on these features, we implement two case studies to show how they could contribute to empirical analysis. We discuss the paper's theoretical contributions and managerial implications, and present our conclusions.

Theoretical Foundations

The visual components of ads play an important role in attracting consumer attention, enhancing ad persuasiveness, and increasing the probability of purchase. The advertising literature contains a large body of work on visual content (Edell and Staelin 1983; Kim and Lennon 2008; Mitchell 1986; Pieters et al. 2010; Pieters et al. 2007; Smith 1991; Tuch et al. 2009; Wu et al. 2016), indicating its superiority over textual content in invoking consumer emotions and recall.

However, there has been little research on visual content in the social media literature. Social media studies on content have tended to focus on author characteristics (Lakkaraju and Ajmera 2011), audience size (Zadeh and Sharda 2014), and textual content (Lee et al. 2018; Ma et al. 2015; Singh et al. 2014; Stieglitz and Dang-Xuan 2013). This paper takes a step forward and incorporates the effect of visual content in social media.

The availability of large and granular social media data has opened up new research avenues. In addition, ML approaches enable us to generate granular measures at scale, which previously required manual coding. However, there is concern in the IS research community that data-driven research could fail to build a cohesive body of knowledge and the conclusions cannot be easily generalized or explained. Therefore, the establishment of solid theoretical foundations is recommended in big data research, since theory can provide guidance in research focus and construct selection (Johnson et al. 2019; Maass et al. 2018; Rai 2016).

We, therefore, extensively review the literature to build a theoretical foundation and to identify visual and textual features that can play an important role in social media ad per-

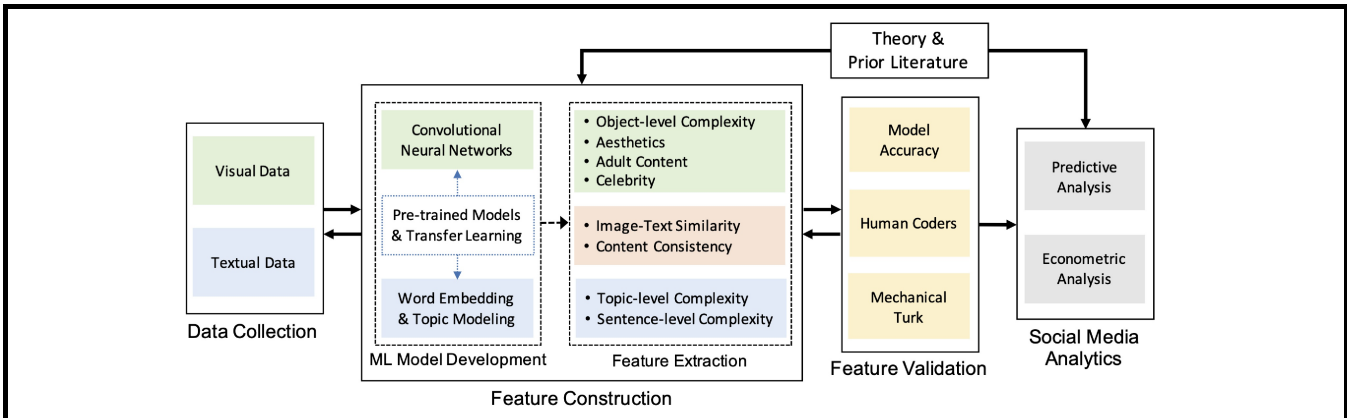


Figure 1. Proposed Research Framework

uasion (theoretically motivated features). In addition, we complement these features with generic visual and textual content characteristics developed from ML models (deep learning-enabled features).

Theoretically Motivated Visual and Textual Features

A stream of studies in the advertising literature has investigated various image-related features as the determinants of ad effectiveness. The potential impact of pictures can be explained with the ELM framework (Petty and Cacioppo 1986; Petty et al. 1983), one of the most popular models of persuasion that provides a framework to understand the effectiveness of persuasive communication. In the ELM framework, two routes explain the complicated process of persuasion: *a central route* and *a peripheral route*. The processing of images can potentially be influenced through both routes. Simple peripheral cues in pictures can attract attention and evoke imagery (MacInnis and Price 1987; Unnava and Burnkrant 1991), while images that are well aligned with the ad can facilitate consumers' elaboration of the information content (Miniard et al. 1991, Pieters et al. 2010). In Table 1, we summarize the visual content features introduced in the advertising literature within the ELM framework. We note that the effectiveness of these visual measures has not been systematically evaluated in the social media setting.

Peripheral Factors

Visual peripheral factors can influence a picture's persuasion effectiveness via the peripheral route. Celebrity or domain expert endorsement is a representative example of such visual cues (Petty and Cacioppo 1986). Social influence theory explains that consumers can be impacted by celebrities due to compliance, identification, or internalization, rather than from the quality of the content (Kelman 1961). Studies showed that celebrity endorsements contribute to consumers' belief in the product's worth and credibility (Agrawal and Kamakura 1995; Friedman and Friedman 1979; Kamins 1989).

Aesthetic stimulus and sexual appeal are also well documented peripheral cues in the literature. Both aesthetics and nudity can induce attention, which can improve ad effectiveness (Severn et al. 1990). Product aesthetics have long been among companies' strategic tools to remain competitive in the market and to improve consumers' product evaluation (Bloch 1995; Page and Herr 2002). Recent studies have also discovered how the aesthetics of web and system design can positively impact consumers' physical and psychological responses (Jiang et al. 2016; Strebe 2016). On the other hand, sexual appeal can detract from consumers' ability to process information in the central route (Steadman 1969).

Finally, pixel-level image complexity (an image's variation at the pixel level) is another widely accepted visual cue that could play a role in the peripheral route. Compared with object-level image complexity, which we will introduce later, pixel-level complexity only evokes low-level visual processes (Pieters et al. 2010). This measure is also referred to as "visual complexity" (Donderi 2006a) or "visual clutter" (Rosenholtz et al. 2007). Studies found mixed results on the effect of pixel-level complexity on ad effectiveness. Ads with an image of high pixel-level complexity can hinder the identification of objects within through the central route, having a negative impact on overall attitudes toward the ad (Donderi and McFadden 2005). On the other hand, pixel-level complexity can also increase physiological arousal, self-reported arousal, and memory (Deng and Poole 2010; Huhmann 2003).

Table 1. Theoretically Motivated Visual Content Features

Feature	Definition	References	Results	Methods
Peripheral Route				
Celebrity endorsement	Celebrity or domain expert endorsement	Agrawal and Kamakura (1995), Friedman and Friedman (1979), Kamins (1989)	Positive impact	Lab experiments with students as subjects, event study on 110 celebrity endorsement contract announcements
Sexual appeals	Visual sexual content, such as nudity, pin-up models, and muscular men	Severn et al. (1990), Steadman (1969)	Mixed impact	Lab experiments with students
Aesthetics	A critical reflection on art, beauty, and taste, with the creation and appreciation of beauty	Bloch (1995), Jiang et al. (2016), Page and Herr (2002), Strebe (2016)	Positive impact	Surveys from experts and students, lab experiments with students
Pixel-level complexity	Images' variation at the pixel level	Deng and Poole (2010), Donderi (2006a), Donderi and McFadden (2005), Huhmann (2003), Rosenholtz et al. (2007)	Mixed impact	Pictures coded by experts, lab experiments with participants
Central Route				
Object-level complexity	Visual structural variation in terms of specific objects, shapes, and arrangements	Deng and Poole (2010), Geissler et al. (2006), Kosslyn (1975), Palmer (1999), Pieters et al. (2010)	Positive impact, U-shaped relationship	Pictures coded by experts, lab experiments with participants, interviews
Content consistency (variety vs. consistency seeking)	The consistency of a specific content with respect to previous contents in the channel	Adomavicius et al. (2015), Fong (2017), Johnson et al. (2006), McAlister (1982)	Mixed impact	Observational analysis on other products
Text-image similarity	The relationship between the contents of text and images	Deng and Poole (2010), Miniard et al. (1991), Phillips (2000)	Mixed impact	Lab experiments with students, picture coded by students

Factors in the Central Route

Visual content can also play an important role during viewers' scrutiny of a picture's content in the central route. One such factor documented in the literature is the "structural complexity" or "design complexity" of images. This feature assesses visual structure variations in terms of specific objects, shapes, and arrangements (Pieters et al. 2010). The higher a picture's design complexity, the more information processing ability it requires of the consumer to understand it. In the print ad context, Palmer (1999) and Pieters et al. (2010) found design complexity to have a positive effect on readers' attitudes toward ads. Other studies on website design showed that an intermediate level of complexity receives the most favorable response (Deng and Poole 2010; Geissler et al. 2006).

Content consistency is another factor that helps a consumer in processing newly introduced information. Social media users often subscribe to specific accounts (e.g., page likes on Facebook or following others on Twitter) to receive continuous information from them. Subscribers' assessment of a specific post from an account can be based on their preferences. Consumers can favor a conventional post if they exhibit stable preferences in social content, namely, *consistency seeking* behavior (Fong 2017; Johnson et al. 2006; Oliver 1999). If consumers are *variety seeking*, however, innovative content can receive more praise (Adomavicius et al. 2015; McAlister 1982; Simonson 1990).

The relations between images and text in a social media post can affect consumer's information processing. In the ELM framework, message repetition can modify receivers' attitudes in a two-stage process (Petty and Cacioppo 1986). In the first stage, textual content that is related to images can enhance consumers' ability to understand the ads. In the second stage, images deliver content in a different format, which can greatly reduce the potential for tedium or reactance. Prior research showed that pictures that are connected to the textual content

of ads tend to receive better customer evaluations (Miniard et al. 1991; Phillips 2000).

Finally, in addition to visual aspects, textual content in social media will affect consumer information processing in the central route, because text comprehension requires significant cognitive effort (Petty and Cacioppo 1986). The psychology literature has shown that textual comprehension can be described in terms of macro-level (global) and micro-level (local) processes: individual words and sentences are processed at the micro-level and the full meaning of the text is organized at the macro-level (Goodman 1967; Gough 1972; Kintsch and van Dijk 1978).

Operationalization in the Literature

We note that the theoretically motivated features mentioned so far were mainly handcrafted in the literature. This operationalization approach has limitations in scalability, since it requires significant manual engineering effort and expert-level domain knowledge. Therefore, the empirical findings of most studies are based on a limited number of picture samples in a relatively small-scale lab setting. For example, Unnava and Burnkrant (1991) conducted experiments with 107 undergraduate students assessing seven ads to make conclusions about the findings. Similarly, Pieters et al. (2010) used a random sample of 249 full-page advertisements evaluated by 100 regular customers and 12 trained judges. Apparently, however, this approach cannot be applied to large-scale social media data with tens of thousands of pictures and the reactions of millions of people. We therefore apply ML approaches to *automate* the construction of theoretically motivated measures, which can address scalability issues and mitigate potential sample selection bias. We note that the theoretically motivated features are *qualitatively* similar to those used in existing literature.

Deep Learning Enabled Features

In addition to the benefit of automating the construction of theoretically motivated features, a deep learning model, specifically a CNN, can also enable qualitatively new visual content features, which we term *generic visual content features*. In the computer vision literature, these generic features, the second-to-last layer in CNN models, have been shown to be, in fact, robust and effective representations of images (called CNN *codes*), which has significantly improved the accuracy of various object detection and, image recognition and classification tasks (Donahue et al. 2014; He et al. 2015; LeCun et al. 2015; Rajpurkar et al. 2018; Razavian et al. 2014; Yosinski et al. 2014; Yu et al. 2016). Although CNN codes have not yet been widely explored in the social media literature, they could influence consumers' evaluation of ads, either consciously or unconsciously.

Similarly, *generic textual content features* can be learned with word embedding deep learning models, which are trained to learn word vector representations that encompass the semantic relations between them (Mikolov et al. 2013). Because of their superior performance and scalability, word embedding approaches have become pervasive in a wide range of natural language processing (NLP) and information retrieval problems, such as sentiment classification (Tang et al. 2014), document classification (Taddy 2015), named entity recognition (Lample et al. 2016), search query completion (Mitra 2015), machine translation (Zhang et al. 2014), and conversational artificial intelligence (AI) systems used in smart assistants such as Amazon Alexa (Ram et al. 2018). Our prediction study (Case Study 1) indeed shows that the inclusion of these generic visual and textual content features can substantially enhance the accuracy of predicting a social media post's popularity.

Visual and Textual Feature Construction

Based on the theoretical foundation, this section describes how we operationalize the visual and textual content features from unstructured data in social media posts. To construct robust representations of image data, we leverage deep learning approaches, including CNNs. For text data comprehension, we use NLP techniques such as topic modeling and word embedding. Based on the textual and visual features, we further construct content consistency and image-text similarity features.

We use Tumblr data as the basis for feature construction.[7] As a social network service, Tumblr's users can follow blogs of interest without mutual confirmation. As a microblogging platform, Tumblr provides a wide range of useful tools similar to those of traditional blogging sites, allowing companies to create long, rich, high-quality content. These tools provide Tumblr's users the flexibility of choosing various designs, unlike Facebook or Twitter, which have a set layout. Therefore, Tumblr blogs closely resemble regular websites, often re flecting the brand's personality. Three examples of companies' official Tumblr blogs with different layouts are shown in Figure 2.

Figure 3 gives an overview of the newly introduced visual and textual content features that are constructed using ML approaches. Typically, a social media post contains visual and textual content. From the image, we introduce object-level complexity, an aesthetics level, an adult content level, and celebrity endorsements. From the text, we propose topic and

[7]We note that feature construction is not tied to Tumblr's data but can be applied to other social media datasets.

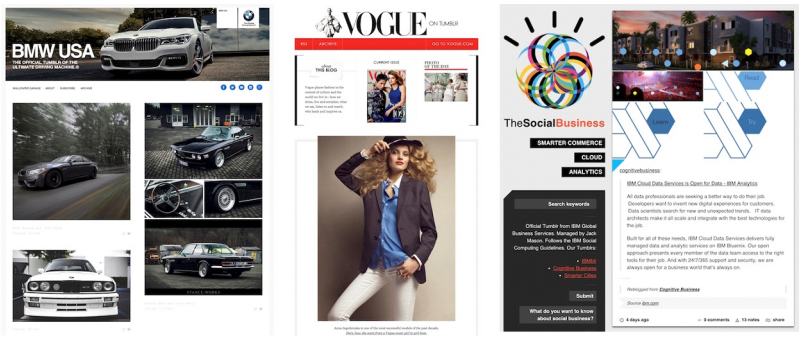


Figure 2. Examples of a Company's Official Tumblr Blogs: BMW, Vogue, and IBM

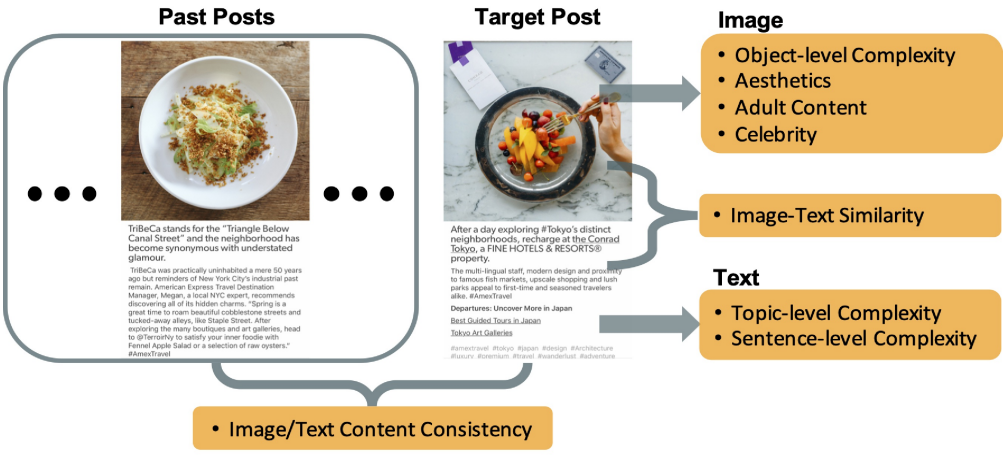


Figure 3. Overview of New Visual and Textual Content Features Leveraging Deep Learning Models (for brevity, we omit other simple content features)

sentence-level complexity measures. We also introduce features capturing relations between content dimensions, that is, image-text similarity within a post and image/text content consistency across posts.

Visual Features

Visual content, such as images, is usually represented as a multidimensional matrix of pixel values, where each pixel represents a low-level data point of the visual stimulus. We construct an image complexity measure based on the low-level pixel data, capturing visual cues that can impact ad effectiveness in the peripheral route within the ELM framework (Petty and Cacioppo 1986; Petty et al. 1983). A complementary approach is to capture the visual content's high-level semantic, which could require consumers to carefully evaluate the message content to achieve a full understanding (Pieters et al. 2010). Deep learning approaches are used to recognize objects in images, to measure the image's aesthetic and adult-content levels, and to detect celebrity endorsements. Object recognition, in turn, is used to construct an object-level complexity measure that could influence consumer's information processing in the central route. In sum, we argue that pixel-level image complexity, the aesthetic, and adult-content levels, and celebrity endorsement can serve as peripheral cues and that object-level image complexity will be an influential factor in the central route.

Pixel-Level Image Complexity

According to visual complexity theory (Attneave 1954; Donderi 2006a), visual stimuli such as images are a composite of different elements, including color, luminance, and texture. An image with more variations in color or brightness will be more complex than one with fewer colors or uniform brightness. As discussed earlier, the impact of pixel-level visual

complexity on persuasiveness can be two-fold, increasing a person's arousal and memory or hindering the information processing required to evaluate the true merit of the message.

An image's compressed file size is widely used to measure such visual complexity (Donderi 2006b; Forsythe et al. 2011, Machado et al. 2015; Pieters et al. 2010; Tuch et al. 2009). It represents the minimal computer storage required to store the image, which increases as variations in one or more such pixel-based features increases. We use an image's normalized compressed file size as *pixel-level image complexity*, since image posts are generated in various resolutions and formats with different compression algorithms (e.g., lossy compression such as JPEG and lossless compression such as PNG and GIF). To obtain a consistent measure of pixel-level complexity across different settings, we normalize the compressed file size f (measured in bytes) of an image by its resolution r (i.e., number of pixels) and compression quality q.[8] We compute pixel-level complexity as $ImagePixelComplexity = (100 \times f)/(q \times r)$.

Object-Level Image Complexity via Deep Learning

While pixel-level complexity effectively captures how visually complex an image is in *appearance*, it does not capture the high-level *semantics* embedded within. We argue that the semantics of visual content can influence consumers' information processing via the central route in the ELM framework.

To analyze the semantics of an image, one needs to detect and classify *objects* that appear in the image, which is an image recognition task in computer vision. Conventional image recognition approaches involved careful feature engineering efforts based on considerable domain knowledge to obtain useful and robust features (Csurka et al. 2004; Lowe 2004), which is why image analysis has relied on basic image features or resorted to manual feature extraction.

Recent breakthroughs of deep learning, such as CNNs, have enabled scalable and accurate methods to detect objects contained in images (LeCun et al. 2015; Krizhevsky et al. 2017). The key aspect of a CNN is that it automatically discovers robust representations needed for accurate classification via the composition of such multiple transformations. In other words, the layers are not designed by humans (which is the case in most traditional methods), but are learned from the data. The CNN model we use in this paper was developed at Yahoo! that drives many of its services, including Flickr (a photo service owned by Yahoo!). Details on CNN models are given in Appendix B.

Images from social media posts are given as input to the trained CNN model to obtain their confidence scores in 1,700 object categories. Then, to measure the object-level content complexity of the images, we employ the Shannon diversity index to measure the variety in the CNN-generated confidence scores for an image. Let $\boldsymbol{p} \in [0, 1]^d$ confidence scores for a given image, where $d = 1{,}700$ in our case. Then, *object-level image complexity* is defined as

$$ImageObjectComplexity = -\sum_{i=1}^{d} p_i \log(p_i) \quad (1)$$

Note that $\sum_{i=1}^{d} p_i = 1$ and the measure obtains a maximum value of $\log(d)$ when $\mathbf{p}$ is uniformly distributed. As the image content becomes focused on fewer object categories, complexity decreases and eventually becomes zero when $p_i = 1$ for some i.

The image examples in Figure 4 highlight the difference between pixel- and object-level visual complexity. Panels (a) and (b) have only a few monotonous colors (e.g., white, dark blue, black), exhibiting low pixel-level complexity, whereas panels (c) and (d) have high pixel-level complexity due to larger variations in their pixel values, which are reflected as more vibrant colors and greater luminance in the images. In terms of object-level complexity, panels (b) and (d) contain many distinct objects, resulting in high object-level complexity, whereas panels (a) and (c) have low object-level complexity, since each is an image of a single object (a bag and a shoe, respectively).

Obviously, the quality of the proposed object-level complexity measure depends on the accuracy of the CNN model. We report that our CNN model achieves 91.9% prediction accuracy on a stratified random sample of 2,500 Tumblr images based on the labels from human coders (see details in Appendix E). In fact, approximately 24% of images from our Tumblr dataset are hosted on Flickr, which is the dataset used to train the CNN model. Finally, in the "Validation of Visual and Textual Features" section, we also validate the proposed object-level image complexity measure, that is derived from the CNN model.

More Visual Content Features from Deep Learning

We now describe how CNN models can be used to construct other image features that can serve as peripheral cues in the ELM framework, namely, an aesthetic score (Bloch 1995; Jiang et al. 2016; Page and Herr 2002; Strebe 2016), sexual

[8] For lossless compression formats, $q = 100$.

(a) Low *ImagePixelComplexity* (0.1) and Low *ImageObjectCompexity* (0.91)

(b) Low *ImagePixelComplexity* (0.14) and High *ImageObjectComplexity* (3.68)

(c) High *ImagePixelComplexity* (1.83) and Low *ImageObjectComplexity* (1.34)

(d) High *ImagePixelComplexity* (1.64) and High *ImageOjbectComplexity* (5.34)

Figure 4. Example Images with Different Pixel-Level and Object-Level Image Complexity Measures

appeal (according to an adult-content score) (Steadman 1969, Severn et al. 1990), and celebrity endorsements (Agrawal and Kamakura 1995; Friedman and Friedman 1979; Kamins 1989).

In the computer vision literature, CNN has been widely adapted to numerous computer vision tasks. The final layer of a CNN is a classifier that uses the previous layer as input and outputs predictions on different objects. The output of the second-to-last layer is considered a fixed-length feature extractor that generates a robust representation of the image, referred to as *CNN codes*. These generic image features have been shown to generalize well to various computer vision tasks and greatly outperform existing methods (Razavian et al. 2014), as discussed earlier. In addition, Yosinski et al. (2014) showed that the parameters of a pre-trained CNN can be transferred and fine-tuned to another model for a different task. This technique, also known as *transfer learning*, is a powerful and common practice in deep learning, especially for smaller datasets, which achieves outstanding performance by taking advantage of pre-trained models that have been trained on much larger datasets. Models trained via transfer learning can generally reach higher accuracy with much less data and computation time than models trained from scratch (Long et al. 2015; Oquab et al. 2014).

For aesthetic scores (Dhar et al. 2011), adult-content scores (Sengamedu et al. 2011), and celebrity detection (Parkhi et al. 2015), we leverage models built at Yahoo! as part of their production vision processing pipeline, which are also used for Flickr (similar to the CNN model). The models use a baseline deep CNN for image classification and fine-tune its parameters using a separate training dataset of more than 100,000 images for each task. Both aesthetic and adult-content scores range from zero to one, where higher scores imply better image quality and more adult-content, respectively. The celebrity detection model can detect more than 450 celebrities in a given image with very high precision. In the "Validation of Visual and Textual Features" section, we validate aesthetic scores against human coders. Image examples of automobiles with different aesthetic scores are shown in Figure 5, where we can observe that panel (b), with a high aesthetic score, has greater light and depth effects than panel (a).

Models for Visual Feature Construction

Training deep learning models with visual data can be challenging in many aspects, such as selection of the right architecture and optimization algorithm. More importantly, they require large amounts of training data labeled by expert human coders and computational resources such as parallel

(a) Low Aesthetic Score (0.14)

(b) High Aesthetic Score (0.58)

Figure 5. Example Images with Different Aesthetic Scores

computing and GPUs. In this paper, we use the Caffe open-source framework for deep learning (Jia et al. 2014) with the aforementioned proprietary CNN models from Yahoo! to construct deep learning-based visual features. Similarly, other major cloud computing companies, including Amazon, Microsoft, and Google, provide API services for common computer vision applications with their own pre-trained models. These services utilize popular open-source deep learning software, such as Keras (Chollet et al. 2015), PyTorch (Paszke et al. 2017), and TensorFlow (Abadi et al. 2015), which provide different capabilities and interfaces.

Alternatively, many pre-trained models that can be immediately utilized with such software libraries have been made publicly available by the deep learning community to share and reproduce academic results. Table 2 gives examples of open-sourced models for various computer vision tasks that could be useful for social media research. However, visual feature construction tasks that are not addressed by existing pre-trained models would require building a new model. Given the complexity of typical deep learning models, this involves gathering large volumes of labeled training data, which can be quite difficult and costly to acquire, and often requires intense computational power. A useful technique to alleviate these issues is transfer learning, described earlier, where the parameters of an open-sourced pre-trained model are further fine-tuned for the problem of interest using the available labeled training data (Long et al. 2015; Oquab et al. 2014; Yosinski et al. 2014). For example, Google Cloud's AutoML service that facilitates building custom ML models provides transfer learning using their pre-trained models. Another well-known and effective technique is data augmentation (Razavian et al. 2014), where minor alterations (e.g., cropping, flipping, or rotations) are applied to the limited training images to generate additional labeled images without actually gathering new labeled images for training the model.

Textual Features

Besides visual content, textual content also conveys important information in social media posts, where the main message is represented by a collection of words. As discussed earlier, text comprehension is described by global and local processes (Kintsch and van Dijk 1978), where the full meaning (topic) of the text is organized at the macro-level and the individual words of a text are processed at the micro-level. Following such theory, we operationalize textual complexity at the macro- and micro-levels. For the macro-level process, we propose a topic-level text complexity measure that captures the diversity of topics covered in the text. Then, for the micro level process of text comprehension, we propose a sentence-level complexity value that measures the required efforts to comprehend individual sentences. To operationalize this measure, we apply a word embedding model to measure the predictability of a sentence and quantify the complexity (or unpredictability) of the individual words in the text.

Topic-Level Text Complexity via Topic Modeling

A simple approach to quantifying a text's topic complexity is to measure how diverse keywords are used in the text. The implicit assumption of this approach is that each unique keyword represents an independent concept. However, some keywords are interrelated (e.g., mobile and phone) or even synonyms (e.g., mobile phone, cell phone, smartphone).

To incorporate the interrelatedness of individual keywords while capturing the topics of the overall text at a global level, we employ the latent Dirichlet allocation (LDA) topic modeling approach (Blei et al. 2003), following its successful applications in the IS literature (Gong et al. 2018; Lee et al. 2016; Lee et al. 2020; Shi et al. 2016; Singh et al. 2014). The underlying assumption of the LDA model is that a document

Table 2. Open-Source Pre-trained Deep Learning Models for Visual Content	
Task	**Reference**
Image Classification & Segmentation	Chatfield et al. (2014), He et al. (2016), Ren et al. (2017), Szegedy et al. (2016)
Face Recognition	Masi, Rawls et al. (2016), Masi, Tran et al. (2016)
Age & Gender Recognition	Levi and Hassner (2015b)
Emotion Recognition	Barsoum et al. (2016), Levi and Hassner (2015a)
Place & Scene Recognition	Zhou et al. (2014)

consists of a small number of latent topics and that the words in the document are the realization of its underlying topics. The LDA produces two outputs: (1) related keywords for each topic and (2) topic distributions for each document (i.e., post). Using a trained LDA model, where we find that Tumblr text is best represented using 20 topics, each text is transformed into a 20-dimensional topic vector. We describe the details on our LDA topic model in Appendix C.

Using document-level topic distributions, we compute each social media post's text complexity at the topic level. Texts covering multiple topics can be considered semantically complex, whereas those concentrating on a single topic can be considered semantically simple. We define the *topic-level text complexity* for each post (similar to object-level image complexity) as the Shannon index, as in Eq. (1). Specifically, **p** in Eq. (1) is set to be the topic distribution of the text for a given post, yielding larger complexity values for more diverse topics. We note that a similar approach was introduced to measure keyword ambiguity in the context of online search advertising (Gong et al. 2018).

Sentence-Level Text Complexity via Word Embedding

For micro-level text comprehension, we propose a *sentence-level* text complexity measure that can capture the text's predictability. In other words, it is supposed to measure how easily a reader can follow (in terms of predictability) each sentence in a blog post. There are existing readability scores in the literature such as the Flesch-Kincaid readability test (Kincaid et al. 1975) and the Gunning fog index (Gunning 1952), which are simple calculations based on the numbers of words, sentences, and syllables in the text.

However, these existing scores might not be directly applicable to social media studies because they do not consider the specific text context and it is difficult for them to capture the unique nature of social media texts, which often include slang, acronyms, and other nonstandard words. Moreover, the text in social media posts is usually short (with an average of 16 words, as shown in Table 5), in which case similar scores are likely to be assigned to many posts but might not reflect the posts' true complexity. Finally, the Flesch score or Gunning fog index are primarily designed for the English language, and are thus not applicable to different languages.

To measure micro-level text complexity considering the unique nature of social media text, we leverage a deep learning approach called *word2vec* word embedding (Mikolov et al. 2013). Specifically, the word2vec model is a neural network that is trained to learn word vector embeddings with the goal of accurately reconstructing the surrounding context of each word. Semantically similar words are thus mapped to nearby points in the learned vector space. The objective of word2vec is to maximize the log-likelihood of the focal word given the surrounding words in each sentence, which enables one to compute the likelihood of each sentence from the learned model.

One of the advantages of this word embedding approach is that it is data-driven. That is, the predictability of a given sentence will differ between different word2vec models trained on different text datasets. This aspect is particularly useful to incorporate the unique nature of social media text. Another advantage is that word2vec can be applied to different languages and even multilingual text, which is often the case in social media. Details on the word2vec model are described in Appendix D.

We train our word2vec model using the text corpus from social media posts. We use $d = 100$ for the dimension of word vectors, which was chosen by cross-validation with respect to the model's accuracy (Mikolov et al. 2013).[9] From the trained word2vec model of Eq. (3) in Appendix D, we can compute the probability p_s of a sentence s in a given post as the pairwise composite log probability:

[9]We note that other reasonable values of the dimension ($100 \le d \le 500$) yield similar empirical results.

Table 3. Text Examples and Their Complexity Scores from a Tumblr Blog

Text body	Topic-Complexity	Sentence-Complexity
You've been slicing, dicing, pitting and peeling your produce all wrong. Until now. Cooking is fun. Eating is fun. Sometimes you want to get the cooking done so you can get to the eating. We can respect that. See all of our cooking hacks that will save you time in the kitchen here.	0.558	0.437
Instant close-up shots of people and objects courtesy of the Lomo'Instant. We know you're already raring to get your hands on the Lomo'Instant, and to help you tide over until it arrives at your doorsteps, we have these test shots to show you straight from the Lomography team in Hong Kong!	1.583	0.820

$$\log p_s = \sum_{i=1}^{T} \sum_{j \neq i, j=i-b}^{i+b} \log p\left(s_j \mid s_i\right)$$

A high p_s value implies that the sentence s is likely to appear based on the neighboring words, and a sentence with a low p_s value would be less expected for the reader in the current context. Thus, we use one minus the average p_s as a post's *sentence-level text complexity*, which can be written

$$TextSentenceComplexity = 1 - \frac{1}{N} \sum_{s=1}^{N} p_s$$

where N is the number of sentences in a given post. Since this text complexity is a newly developed measure, we validate it in the "Validation of Visual and Textual Features" section, where we also show that it aligns better with human coders than the Flesch readability score does.

In our case studies, we employ both LDA and word2vec to analyze textual information at the topic and sentence levels. Table 3 shows examples of text bodies from two different posts. The first example focuses on a single topic (cooking), resulting in lower topic-level complexity compared to the other example, which includes words that can be related to different topics. The first example also has lower sentence-level complexity, since it consists of short, straightforward sentences. In contrast, the second example contains a long compound sentence, resulting in high sentence-level complexity.

Models for Textual Feature Construction

For both the LDA and word2vec models, we use implementations from the GENSIM package (Řehůřek and Sojka 2010) and the models are trained using Tumblr text data. These models are unsupervised and thus do not require labeled training data. Furthermore, many fast and accurate algorithms for NLP tasks have been recently developed, making model training relatively accessible, and are available in many deep learning software packages mentioned earlier (Bojanowski et al. 2017; Pennington et al. 2014). Nonetheless, open-sourced pre-trained word embedding vectors can be quite useful, especially in cases where the training data are sparse (e.g., languages other than English) (Devlin et al. 2019; Grave et al. 2018; Heinzerling and Strube 2018). Transfer learning can also be applied to extend to other NLP tasks, such as sentiment analysis and question answering using pre-trained word embedding models analogous to the CNN transfer learning described earlier.

Features Capturing Content Relations

We further propose novel features that capture the content relations within a post and between different posts by utilizing the aforementioned deep learning techniques. Specifically, we measure the relatedness of an image and text for a given post with *image-text similarity* and the similarity of a focal post to previous posts using *content consistency*. These feature constructions are possible due to the ability to use a robust and common representation for different types of content.

Image-Text Similarity

A post's visual content and textual content serve different roles but are presented to consumers simultaneously as a single unit (Petty and Cacioppo 1986). Therefore, the interaction between these two distinct types of content is an important factor in terms of how consumers engage with social media posts. Intuitively, a coherent post should consist of images and words that can be easily associated with each other. Either an image should illustrate the story of the text or the text should relate to the image. However, quantifying the relationship between pixel-based images and character-based text is not a straightforward task.

Here, we propose a novel *image-text similarity* measure with the aid of ML methods for image and text analyses. To measure the similarity of two different content types, they need to be transformed into a common representation, which is possible through deep learning approaches. Specifically, we represent each image as a collection of the predicted labels obtained from our deep CNN models and construct a separate "image label corpus" using such representation. In addition to the CNN model, we utilize the celebrity detection model to incorporate celebrity information. Note that there is no natural ordering of words in the image corpus unlike words in regular sentences.

From the combined data of the text and image label corpora, we build an LDA topic model with 50 topics and obtain the topic distributions of both the image and text.[10] We note that tags are added to the text corpus to better capture the relation between the image and textual content, since tags as keywords can also be related to the image. Finally, we measure the content similarity between the image and the text of a given post (*Image-Text Similarity*) as the cosine similarity of the two corresponding topic distributions $\boldsymbol{p}_{image}$ and $\boldsymbol{p}_{text}$:

$$ImageTextSimilarity = \frac{\boldsymbol{p}_{image}^{T} \cdot \boldsymbol{p}_{text}}{\| \boldsymbol{p}_{image} \| \cdot \| \boldsymbol{p}_{text} \|} \quad (2)$$

Table 4 shows three post examples and their image-text similarity, where the first example has high similarity due to the many overlapping words between the image's predicted labels and text. The second example has moderate similarity because the image's predicted labels do not include certain words in the text because they are not highlighted in the image ("nail," "ring") or are difficult to infer from the image ("weekend," "Sunday"). The third example in Table 4 has zero similarity since the text relates to the image in an indirect manner.

Content Consistency

In social media platforms, a steady and continuous readership is formed based on *following* behavior. Content consistency is an important factor that can help a consumer process newly introduced content from channels to which the consumer has subscribed (Fong 2017; Johnson et al. 2006; Oliver 1999). For this, we develop a content consistency measure, that evaluates whether an individual post is similar to or distinct from the usual or average content of a blog. The variable is based on the proposed visual and textual content measures described in previous sections.

Specifically, for post i of a given blog, we compute the average content as $c_i^{avg} = \sum_{j \in \Omega_i} c_j \Big/ |\Omega_i|$, where Ω_i is the set of posts created by the blog prior to post i. For images, c_i is set to be the predicted labels obtained from the CNN model, as discussed earlier. We emphasize that the average image content would be difficult to compute without the representation obtained by a CNN, since images have various resolutions and formats. For text, we set c_i as the corresponding topic distribution computed via the LDA, as discussed earlier. Finally, we measure the *content consistency* of post i as the cosine similarity between c_i and c_i^{avg} similar to Eq. (2). From this, we obtain consistency measures for both the text (*TextConsistency*) and images (*ImageConsistency*). We note that the topic-based text similarity measure was well-adopted in the IS research (Lee et al. 2016; Shi et al. 2016).

Figure 6 illustrates the cumulative distribution function of content consistency over all posts for both text and images. About 45% of texts have a consistency value of 0.6 or greater, whereas only 12% of images reach such a value. That is, company blogs tend to use text with similar topics but adopt images with more diverse objects in their posts. Since the consistency measure is a straightforward calculation from the LDA and CNN models, we do not carry out a separate validation.

Variable Construction

Our Tumblr dataset consists of 34,558 posts created by a panel of 180 official company blogs from various industry sectors over a six-month period between May 2014 and October 2014. The list of companies is given in Appendix A. Among the collected posts, 88.4% are photo posts with text, 7.4% are pure text posts, and the remaining 4.2% are video posts. A total of 53,417 images were collected from all photo posts. The dominance of photo posts exemplifies the importance of our study on visual data analytics. For each post, our data also contain two kinds of consumer engagement measures, the numbers of likes and reblogs, which are collected through April 2015. We supplement these data with posts' visual and textual features that are created by ML algorithms.

Table 5 summarizes the variables used in the analysis and their descriptive statistics. We observe that the distributions of the numbers of reblogs and likes are skewed. On average, a post receives 451 reblogs and 535 likes.

[10] As discussed earlier, the number of topics is determined by consulting multiple criteria, as shown in Figure C1(b), which suggests that 50 to 70 topics is a good choice.

Table 4. Example Posts and Their Image–Text Similarity Scores

Image			
Deep CNN output	quilt, comforter, comfort, couch, bed, window, bedroom, dorm, room, home, living room, bedspread, sofa, bed sheet, vintage, furniture, headboard	coffee mug, espresso cup, classroom, coffee shop, coffee, food, indoor, writing, beverage, hand, tea	jean, blue jean, denim, sweatshirt, closet, clothing store, craft, handkerchief
Text & Tags	Dream bed; interior, design, bedroom, bedding, comforter, apartment, dorm, home	Sunday; mani, coffee, rings, white nail polish, mug, laptop, Sunday, weekend	Find out how Forage Haberdashery and their bowties came into existence; dreamers, doers, bowtie, cloth, Forage Haberdashery
Image-Text Similarity	0.768	0.342	0

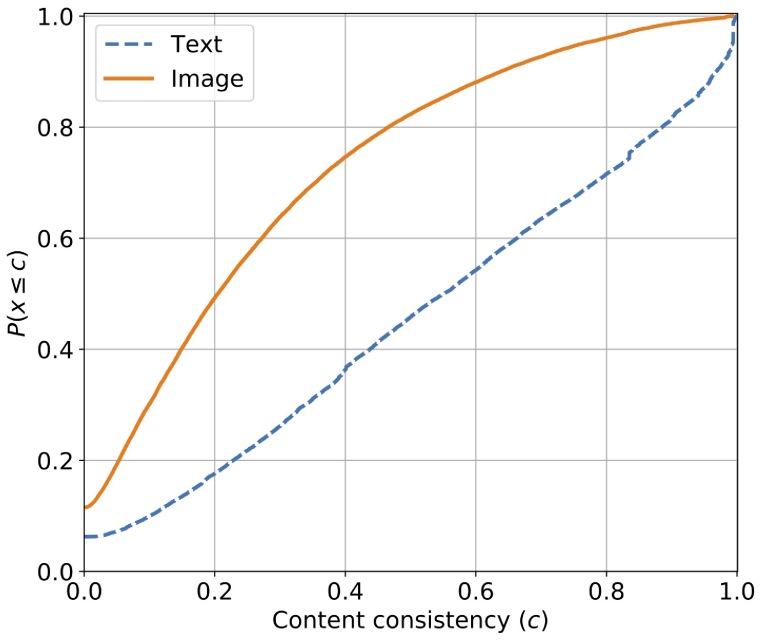


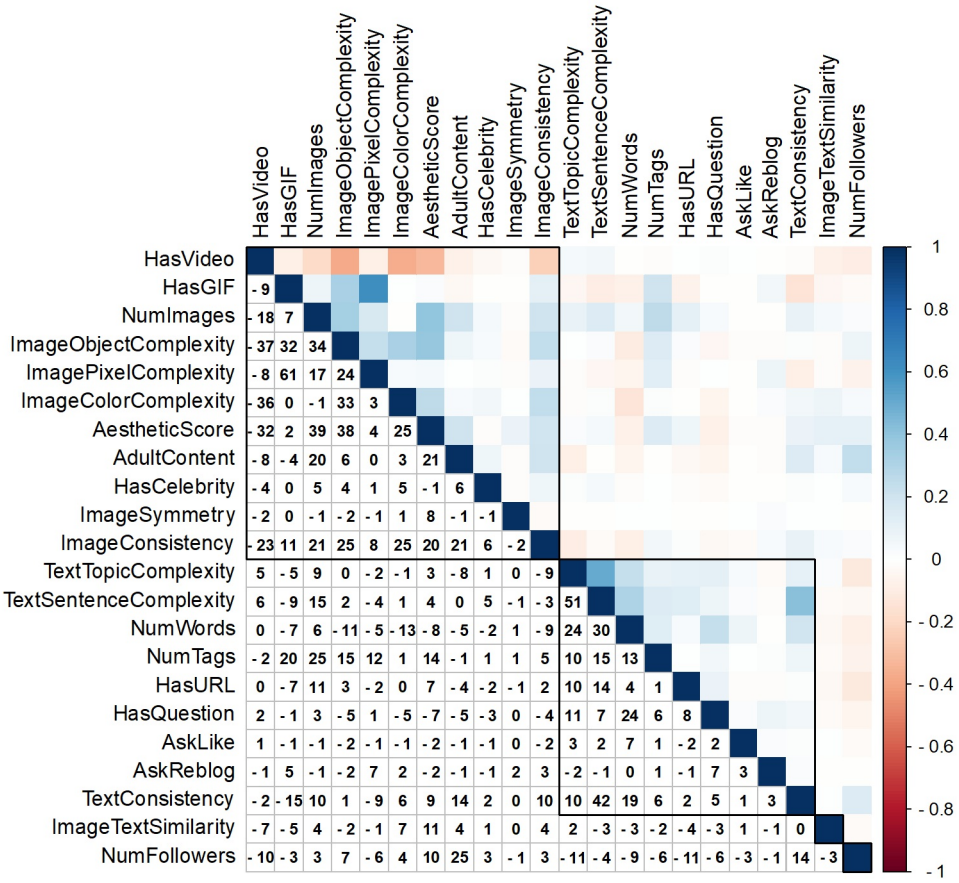
Figure 6. Plot of Text and Image Content Consistency of All Posts Measured by the Cosine Similarity Between the Focal Post and the Corresponding Blog's Average Content

Table 5. Descriptive Statistics of Dependent and Independent Variables

	Mean	Standard Deviation	Min	Max
Dependent Variables				
Likes	534.9	6,727	0	430,745
Followers	36.91	140.9	0	12,624
Non-Followers	498.0	6,667	0	428,009
Reblogs	451.2	6,988	0	521,709
Followers	28.45	166.1	0	17,224
Non-Followers	422.8	6,919	0	518,650
Independent Variables				
Visual				
HasVideo	0.0415	0.199	0	1
HasGIF	0.156	0.363	0	1
NumImages	1.488	1.685	0	10
*ImageObjectComplexity**	0	1	-1.789	2.041
*ImagePixelComplexity**	0	1	-0.391	25.138
*ImageColorComplexity**	0	1	-1.735	1.982
*AestheticScore**	0	1	-1.561	3.024
AdultContent	0.0568	0.153	0	0.995
HasCelebrity	0.0341	0.181	0	1
ImageSymmetry	0.00648	0.0802	0	1
*ImageConsistency**	0	1	-1.125	3.045
Textual				
*TextTopicComplexity**	0	1	-1.218	5.114
*TextSentenceComplexity**	0	1	-2.262	2.443
LogNumWords	2.778	1.198	0	8.419
LogNumTags	1.899	0.66	0	3.434
HasURL	0.111	0.314	0	1
HasQuestion	0.104	0.305	0	1
AskLike	0.00414	0.0642	0	1
AskReblog	0.000984	0.0314	0	1
*TextConsistency**	0	1	-1.735	1.465
Visual & Textual				
ImageTextSimilarity	0.0385	0.113	0	1
Other				
LogNumFollowers	5,692	10,453	0	78,704
Total observations: 34,558	(*indicates that the variable is standardized)			

Since our newly introduced visual features have different scales, we standardize these visual content features (mean = 0, standard deviation = 1) including *ImageObjectComplexity*, *ImagePixelComplexity*, *AestheticScore*, and *ImageConsistency*. In addition, we construct basic features that can be measured without an ML approach. The binary variables *HasVideo* and *HasGIF* depend on the existence of the corresponding media type in the post. The variable *NumImages* represents the number of images in a post, which can have up to 10 images. We also include color complexity (*ImageColor Complexity*), which is computed by Eq. (1) using the color distribution of an image and then standardizing (Pieters et al. 2010).[11] The variable *ImageSymmetry* indicates whether an an image is mostly symmetric (Attneave 1954).

[11]The colors are mapped to their closest color in a standard 16-color palette.



Note: Colors in the upper triangular part represent correlations and the lower triangular part shows the same correlations in percentages (between -100 and +100).

Figure 7. Correlation of Independent Variables

For text-related features, we standardize our text content features, including *TextTopicComplexity*, *TextSentenceComplexity*, and *TextConsistency*, as we did with the visual content features. In addition, we include binary variables indicating the presence of links (*HasURL*) or questions (*HasQuestion*) in the text. Explicit solicitations for likes and reblogs in the text, such as "Like/Reblog if ..." are controlled using corresponding binary variables (*AskLike and AskReblog*).

Figure 7 gives the matrix of correlations between the features. The largest correlation, 0.61, is observed between *HasGIF* and *ImagePixelComplexity*, since GIFs are animated images that usually require more computer storage. The variable *TextSentenceComplexity* has a positive correlation with *TextTopicComplexity*, since sentences consisting of frequent keywords in a given topic are expected to result in a high probability. The variance inflation factor (VIF) is 1.46 ensuring there are no multicollinearity issues with the dataset.

Validation of Visual and Textual Features

Although recent studies show that a deep learning approach can achieve or even surpass human-level performance in image recognition and classification tasks (He et al. 2015; Rajpurkar et al. 2018; Yu et al. 2016), whether the proposed visual and textual content measures actually reflect human perceptions of the measure has yet to be verified. In this section, we validate five measures—object-level image complexity, image aesthetic scores, topic-level text complexity,

sentence-level text complexity, and image-text similarity—using human coders from AMT.[12]

The main goal of the proposed visual and textual measures, rather than employing simple quantitative measures (e.g., image compressed file size, pixel, or word count), is to systematically quantify conceptual aspects of images and text that can be used in empirical analyses. For such abstract measures, the value itself is usually hard to interpret and an intuitive explanation is difficult to obtain from a particular value. For example, an object-level complexity of 1.5 for an image does not necessarily imply that the image is five times more complex than one with a complexity of 0.3. What matters is that the magnitude of the measure provides a means of *comparing* conceptual aspects between different images (e.g., the former image is more complex at the object level than the latter image).

Furthermore, given a single image or text, asking for a score (e.g., 1 to 5) of how an abstract concept is reflected in it can be extremely difficult to answer in an objective and consistent manner. Questions would be in the lines of "How complex do you think the image is?" or "How coherent do you think the post (between image and text) is?" These types of questions can be severely affected by the individual's subjectivity and difficult for human coders to maintain consistency in their answers, especially as they progress through the questionnaire encountering new images and text that could change their relative scale of the measure. In essence, we are interested in whether the proposed visual and textual content measures newly introduced in this paper could successfully produce a *ranked order* of images or text similar to how humans would perceive and order them according to the concept of interest.

Thus, in the validation, we test whether the ML-generated measures can rank the images and text similarly to how humans would, according to the measure of interest. We construct a stratified random sample of 6,000 pairs of posts. About 700 distinct workers participated in our AMT experiment, where each pair of posts was assigned to at least five workers. For all randomly chosen pairs, we asked workers to select an option that better reflects each of the five target image and text measures. In such way, workers have a point of reference to compare against with the target concept in mind. The Cronbach's α (a measure of the internal consistency or reliability of a set of test items) of the aggregated AMT results is 0.8, which exceeds the commonly accepted threshold of 0.7. We follow a number of known best practices for AMT tasks suggested in the literature such as ensuring human worker quality and post-processing AMT results (Lee et al. 2018). The details of the AMT survey instrument are described in Appendix F.

The option that received the majority vote from the workers is considered as the final selected option of each pair. The percentage of pairs that match the order given by the workers (also known as the Kendall τ distance) is used to evaluate how the proposed measure agrees with human interpretation. Specifically, we examine the match percentage with respect to the absolute difference $\delta_i = \left| m_i^{(A)} - m_i^{(B)} \right|$, where $m_i^{(A)}$ and $m_i^{(B)}$ are values of the targeted measure for options A and B of pair i, respectively ($1 \le i \le 6000$).

The main results of the AMT surveys are summarized in Figure 8, where we show the match percentages of pairs whose δ_i is larger than the 0% (i.e., all pairs), 25%, 50% (or median), and 75% quantile of all δ_i's. Overall, we observe that all five measures received about 69%–74% match when we use all AMT labels regardless of the level of agreement among the human coders. This is comparable to the ranking correlation results from state-of-the-art studies in the social media literature (Lv et al. 2017; Wang and Zhang 2017). Interestingly, as we progressively increase the threshold from the 25% to 75% quantile and focus only on pairs with large δ_i (more distinctive pairs according to our measures), the match percentage increases to 82%–88%. This shows that ordering of a pair in terms of the target measure is more obvious to human coders when its absolute difference δ_i between the options is larger.

In Table 6, we consider only post pairs that received unanimous votes from human coders. The match percentages of our measures increase to 81%–93% in this subsample, demonstrating that they can accurately reconstruct the relative order of posts if the distinction is clear from the majority of human coders. We also find that only 29%–53% of the post pairs received unanimous votes on their relative ordering from the human coders in Table 6, which shows that the measurement of visual and textual features is not a trivial task. Additionally, we compare the average absolute difference of pairs where at least one worker disagreed (= $\overline{\delta}_D$) and to that of pairs that received a unanimous vote (= $\overline{\delta}_U$) by calculating the increase percentage $\rho = \left(\overline{\delta}_U - \overline{\delta}_D\right) \Big/ \overline{\delta}_D$. In Figure 9, we can observe that there is a 29%–45% increase from $\overline{\delta}_D$ to $\overline{\delta}_U$, illustrating that pairs with a unanimous vote have a much larger difference δ_i in the target measure than pairs with a

[12]We note that other measures (pixel-level image complexity, adult-content scores, celebrity detection, content consistency) were already validated in previous studies cited in the previous section.

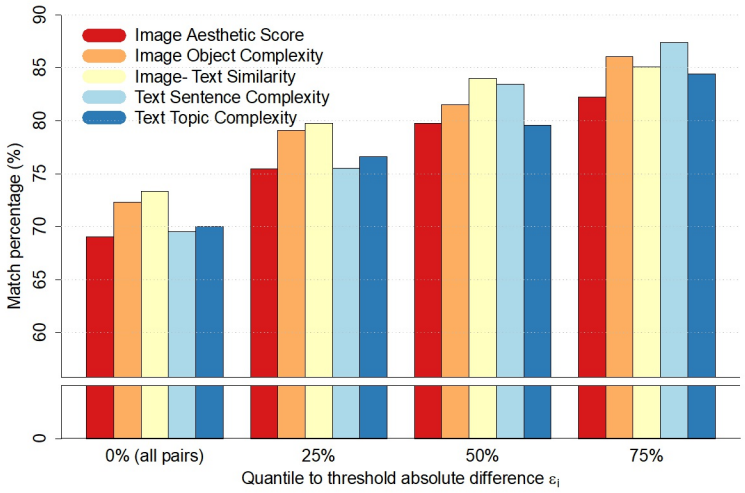


Figure 8. Match Percentage of Pairs by Increasing Quantile Thresholds

Table 6. Results of Pairs with a Unanimous Vote

Measure	Match Percentage	Percentage of Pairs
Image Aesthetic Score	89.7%	34.6%
Image Object Complexity	92.9%	31.3%
Image-Text Similarity	81.6%	32.2%
Text Sentence Complexity	86.4%	29.6%
Text Topic Complexity	90.5%	53.0%

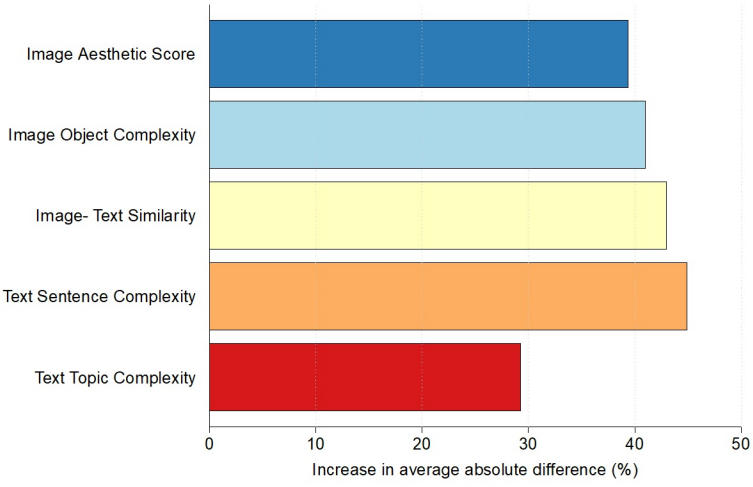


Figure 9. Increase in Average Absolute Difference (ρ) from Pairs with a Divided Vote ($\overline{\delta}_D$) to Pairs with a Unanimous Vote ($\overline{\delta}_U$)

divided vote.[13] These validation results support that the proposed measures align with human interpretation.

Finally, as a comparison to our *TextSentenceComplexity*, we report that the existing Flesch readability score achieves a 66% match with the AMT results, which increases up to 81% at the 75% quantile threshold for δ_i.[14] This is lower than the 69%–88% match achieved by *TextSentenceComplexity* as shown in Figure 8, where the gap increases as the threshold quantile increases. The increase percentage ρ is 45% for the *TextSentenceComplexity* as shown in Figure 9, while the Flesch score gives a much smaller 26% increase percentage. The comparison shows that *TextSentenceComplexity* better aligns with human coders on social media text readability than the traditional readability score.

Case Studies on Visual Data Analytics

To empirically show the effectiveness of our proposed visual data analytics framework, we conduct two case studies: a predictive analytics study and an explanatory econometric modeling (Lee et al. 2020; Shmueli and Koppius 2011).

Case Study 1: Social Media Popularity Prediction

Social media popularity prediction has widespread applications, such as ad placement, online marketing, and trend detection, and has thus attracted extensive attention in the computer science literature (Bandari et al. 2012; Fontanini et al. 2016; Gelli et al. 2015; Lakkaraju and Ajmera 2011; Lerman and Hogg 2010; Wu et al. 2017; Zadeh and Sharda 2014). Most of these previous works, however, focused on particular aspects of social media posts (e.g., audience size, textual features) without much theoretical support.

Our first case study is to assess the predictive power of the newly proposed visual and textual features constructed from our proposed framework. The goal is to accurately predict the popularity of a post, that is, whether it will go viral or not. We grouped the visual and textual features according to the following guiding principles: (1) Is the focal feature constructed without deep learning? (2) Does the focal feature already exist in the literature and the construction can be *automated* with deep learning? (3) Is the focal feature only operationalizable with the use of deep learning approaches? Specifically, we compare the prediction accuracy of three different sets of features including (1) baseline features extracted without deep learning approaches used in prior studies, such as *LogNumFollowers*, *ImagePixelComplexity*, and *LogNumWords*; (2) theoretically motivated features that can now be algorithmically derived from deep learning methods including *AdultContent*, *AestheticScore*, *HasCelebrity*, *ImageObjectComplexity*, *TextSentenceComplexity*, *ImageConsistency*, and *ImageTextSimilarity*;[15] and (3) deep learning-enabled visual and textual features, which are generic representations of images and text learned by the CNN and word2vec models, respectively, as discussed earlier. The deep learning-enabled visual features are the 4,096-dimensional CNN code vectors obtained from the second-to-last layer of the CNN model (see Appendix B), and the corresponding textual features are 100-dimensional word vectors from the word2vec model, as described earlier.

We evaluate the impact of these features on prediction accuracy as we incrementally combine the three feature sets as shown in Table 7. Comparing F1 and F2 will show if the theoretically-driven features can improve prediction performance. Although the deep learning-enabled features are not directly interpretable, they have been shown to produce the best results in many ML applications (Donahue et al. 2014; He et al. 2015; Rajpurkar et al. 2018; Razavian et al. 2014; Yu et al. 2016). Comparing F2 and F3 will reveal whether the deep learning-enabled generic features contain useful information for prediction that is not captured by theoretically motivated features in F2. We note that F2 is the same dataset we use in the following subsection for our second case study, which is described in Table 5. As a measure of post popularity, we use both the numbers of likes and reblogs a post received. Since the numbers of likes and reblogs follow power-law distributions, we apply a log transformation to make them resemble Gaussian distributions.

For each case, the dataset (features and post popularity measures) is randomly split into two, where the first part is used for training and the other part for testing. We report the root mean squared error (RMSE) of the test set averaged over 50 different random splits of the dataset. Additionally, we treat post popularity prediction as a *ranking* problem and measure how well the prediction models can recover post rankings in terms of popularity. Specifically, we compare the popularity

[13] T-test indicates a significant difference between $\bar{\delta}_U$ and $\bar{\delta}_D$ at the 0.01-level for all target measures.

[14] The Flesch readability score assigns lower numbers to text that are more difficult to read. Thus, we take the negative of the scores and normalize to [0, 1] to match the scale with *TextSentenceComplexity*.

[15] Note that *TextTopicComplexity* and *TextConsistency* are part of the baseline features, since they can be constructed without a deep learning approach.

Table 7. Description of Datasets Used in the Prediction Model

Dataset	Description (Dimension)	Features
F1	Baseline Features (15)	*HasVideo, HasGIF, NumImages, ImagePixelComplexity, ImageColorComplexity, ImageSymmetry, TextTopicComplexity, LogNumWords, LogNumTags, HasURL, HasQuestion, AskLike, AskReblog, TextConsistency, LogNumFollowers*
F2	Baseline + Theoretically Motivated Features (22)	Baseline Features + *AdultContent, AestheticScore, ImageConsistency, HasCelebrity, ImageTextSimilarity, ImageObjectComplexity, TextSentenceComplexity*
F3	Baseline + Theoretically Motivated + Deep Learning Enabled Features (4,218)	Baseline + Theoretically Motivated + Generic CNN features (4,096) + Generic word2vec features (100)

Table 8. Comparisons of RMSE for Different Feature Sets Across Different Models

Model	Likes			Reblogs		
	F1	F2	F3	F1	F2	F3
kNN	0.658	0.624	**0.559**	0.763	0.728	**0.666**
Lasso	0.624	0.620	**0.557**	0.723	0.718	**0.648**
SVR	0.589	0.552	**0.499**	0.658	0.654	**0.647**
RF	0.569	0.558	**0.488**	0.686	0.685	**0.647**
FFNN	0.589	0.545	**0.486**	0.664	0.651	**0.610**

Table 9. Comparisons of SRC for Different Feature Sets Across Different Models

Model	Likes			Reblogs		
	F1	F2	F3	F1	F2	F3
kNN	0.559	0.633	**0.723**	0.528	0.583	**0.668**
Lasso	0.668	0.670	**0.748**	0.613	0.621	**0.708**
SVR	0.715	0.729	**0.799**	0.680	0.684	**0.720**
RF	0.713	0.724	**0.776**	0.639	0.640	**0.707**
FFNN	0.700	0.739	**0.808**	0.670	0.685	**0.751**

rankings according to the predicted scores and ground truth using the Spearman ranking correlation (SRC), which was used as the main evaluation metric in recent social media prediction challenges.[16] The SRC range is in [−1, 1], where a score of one corresponds to perfect correlation between the predicted and actual rankings, whereas −1 corresponds to an inverse correlation between the two rankings. We test the three feature sets on a variety of different ML models, including k-nearest neighbor (kNN), Lasso, support vector regression (SVR), random forest (RF), and feed-forward neural network (FFNN). All features are standardized before training and the hyperparameters of the models are tuned via five-fold cross-validation on the training set.

Tables 8 and 9 report the comparisons of the RMSE and SRC, respectively, for different feature sets across the ML models. The best performance is given by F3 in terms of both the RMSE and SRC, and the use of F1 alone has the worst accuracy. We can also see that moving to F3 gives a significant boost in performance, compared to the incremental gain from F1 to F2, illustrating the effectiveness of the deep learning-enabled features. For example, Table 8 shows 7.5% and 17.5% reductions in the RMSE to predict the numbers of likes with an FFNN using F2 (0.545) and F3 (0.486) compared to F1 (0.589), respectively. In terms of SRC, we note an increase of 5.6% for F2 (0.739) and 15.4% for F3 (0.808) compared to F1 (0.700), as shown in Table 9. Furthermore, such a performance trend can be observed in all the tested maximum likelihood models.[17] We note that including the

[16]Social Media Prediction Challenges at the ACM Conference on Multimedia (http://www.acmmm.org/2017/challenge/social-media-prediction/).

[17]The results of t-tests indicate significant differences between comparisons at 1% level of significance.

Table 10. Top-k Importance Ranking of Features by Random Forest

Dataset		k	Feature Importance Ranking
Likes	F1	10	*LogNumFollowers, ImagePixelComplexity, LogNumTags, NumImages, HasGIF, LogNumWords, TextConsistency, HasQuestion, ImageColorComplexity, TextTopicComplexity*
	F2	17	*LogNumFollowers, ImagePixelComplexity, LogNumTags, NumImages, AestheticScore, ImageConsistency, ImageObjectComplexity, HasGIF, ImageTextSimilarity, AdultContent, TextConsistency, LogNumWords, TextSentenceComplexity, ImageColorComplexity, HasCelebrity, TextTopicComplexity, HasQuestion*
	F3	54	*LogNumFollowers, LogNumTags, ImagePixelComplexity, NumImages, AestheticScore, ImageConsistency(6), · · · , ImageObjectComplexity(8), · · · , HasGIF(11), · · · , AdultContent(23), ⋯ , ImageTextSimilarity(35), ⋯*
Reblogs	F1	10	*LogNumFollowers, LogNumTags, ImagePixelComplexity, NumImages, LogNumWords, HasGIF, TextConsistency, ImageColorComplexity, TextTopicComplexity, HasQuestion*
	F2	15	*LogNumFollowers, LogNumTags, ImagePixelComplexity, NumImages, ImageTextSimilarity, HasGIF, AestheticScore, ImageObjectComplexity, ImageConsistency, TextTopicComplexity, LogNumWords, TextConsistency, ImageColorComplexity, HasQuestion, TextTopicComplexity*
	F3	80	*LogNumFollowers, LogNumTags, ImagePixelComplexity, NumImages, HasGIF(5), · · · , AestheticScore(8), · · · , ImageObjectComplexity(11), ImageTextSimilarity(12), ⋯ , ImageConsistency(17), ⋯ , TextSentenceComplexity(19), ⋯ , LogNumWords(21), · · · , HasQuestion(78), · · ·*

text topic vectors in F3 does not have much of an impact on the results.

Among the different models, FFNNs yield the best performance, closely followed by the SVR and RF models. We note that an FFNN is a deep learning approach based on neural networks. The unique aspect of this approach is that the model can learn important high-order interactions between the input features. In other words, an FFNN can learn features that are not captured by the feature engineering efforts made with domain expertise. Thus, this prediction experiment shows that deep learning approaches can enhance social media prediction accuracy in both the feature generation stage (visual and textual content features) and the model learning stage.

The results so far show us that theoretically motivated and deep learning-enabled feature sets make significant contribution in boosting prediction accuracy. Next, we want to determine the importance of individual features. Table 10 shows the feature rankings in terms of their importance for prediction, as determined by the RF model. For F2, we observe that many theoretically motivated features derived from deep learning models, such as *ImageTextSimilarity*, *AestheticScore*, *ImageObjectComplexity* and *ImageConsistency*, are selected as important features for social media popularity prediction. For F3, many of the deep learning-enabled generic features emerge are in the top-ranked list. Note that the theoretically motivated features still appear in the list, implying that theoretically-driven features are useful even in the presence of generic features. In all cases, the results show that *LogNumFollowers*, *LogNumTags* and *ImagePixelComplexity* appear as the most important features. Lastly, we note that the selected features are consistent across the results with likes and reblogs and that Lasso regression also selected similar feature subsets as the RF model, indicating the robustness of our findings.

Case Study 2: Determinants of Social Media Ad Effectiveness

In the second case study, we conduct an econometric analysis of the impact of the proposed visual and textual content features on consumer engagement and measure how the features affect consumer engagement in terms of the numbers of likes and reblogs.

We follow the arguments from the "Theoretical Foundations" section and use the ELM framework to analyze the impact of visual content on the persuasiveness of social media ads (Petty and Cacioppo 1986; Petty et al. 1983). Since social media platforms have become extremely crowded, social media users are facing *massive information overload*.[18]

[18]On average, a user will receive 1,500 posts on their Facebook news feed each day. Source: Social Pilot, "217 Social Media Marketing Statistics to Prep You For 2019" (vailable at https://goo.gl/kpSjPz).

Therefore, users tend to have limited attention levels to thoroughly use their cognitive ability in processing and evaluating the quality of an individual ad's content (Pieters et al. 2010, Teixeira et al. 2012). Thus, social media users could lack strong motivations to scrutinize the meaning of a complicated social media post. Considering these facts, we predict that the effect of peripheral cues will be pronounced in the social media context. Hence, we expect that ads containing images with positive peripheral cues (e.g., pixel-level image complexity, aesthetic pictures, adult content, celebrity endorsements) will be more persuasive for consumers, leading to greater consumer engagement.

Conversely, people are less likely to put cognitive effort into processing ads with complicated content via the central route. Therefore, visual and textual content that require less ability to determine the value (Petty and Cacioppo 1986), such as repeated and consistent content, can improve viewers' attitudes about a post (Lane 2000; Lien 2001). In this case study, we explore if this finding will also hold in the social media context. Specifically, companies often explain their content in detail by adding one or more pictures with supporting textual descriptions. If the visual part of a message is closely related to its text, it can potentially facilitate consumers' understanding of the message, since its recipients are given additional sources with which to process the information. Similarly, individual posts that are consistent with the blog's average image and topics, with which the blog's followers are quite familiar, will be more likely to be positively perceived by the followers.

In our econometric model, the dependent variables are the numbers of likes and reblogs that a Tumblr post receives, which are count variables with over-dispersion. We use fixed effects negative binomial regressions (Nbreg) for the main analysis to control for blog-level, time-invariant unobserved characteristics. We also include linear regression with logarithmically transformed numbers of likes and reblogs as a robustness check.[19]

Table 11 summarizes the empirical results. Regarding the effect of visual peripheral cues (e.g., an image's pixel-level complexity, aesthetic score, adult-content score, and celebrity endorsements), the results show that these factors have positive effects on consumer engagement. The results are qualitatively consistent across different specifications. Specifically, a one standard deviation increase in an image's pixel-level complexity measure (*ImagePixelComplexity*) will result in the post receiving about 6%–12% more reblogs and about 4%–10% more likes. On the other hand, the coefficients of the visual and textual features that increase the required cognitive efforts to process the message are mostly negative and statistically significant. For the visual features, images with higher object-level complexity (*ImageObjectComplexity*) tend to have fewer likes and reblogs: a one standard deviation increase in the measure will induce a decrease of about 5% and 7% in the numbers of reblogs and likes. In terms of the text complexity measure (*TextTopicComplexity* and *TextSentenceComplexity*), the estimated coefficients of *TextSentenceComplexity* are also negative and significant, indicating that sentence-level complexity plays a more pronounced role than topic-level complexity in people's information processing on social media. Furthermore, we can see that the coefficients of the similarity between the image and the text (*Image-TextSimilarity*) and a post's visual and textual content consistency (*ImageConsistency* and *TextConsistency*) are all positive and statistically significant for the majority of the specifications. The results demonstrate the beneficial impact of repetition and consistency in social media ad persuasion. To summarize, the results show that social media users prefer posts with simpler content that requires fewer cognitive resources.

[19]Poisson regression is not applicable in our setting because the distributions of the dependent variables do not meet Poisson model's assumption.

Contributions and Implications

In this section, we discuss the contributions of this research methods article. We first articulate the methodological contributions to the IS and related literature. Then we consider managerial and societal implications of the proposed visual data analytics framework.

Contributions to the Literature

A significant portion of big data is in unstructured data formats, such as text, images, and videos. However, to our knowledge, few IS and social media studies have incorporated large-scale visual content analysis. ML and NLP techniques can help analyze such unstructured data in extracting patterns and insights. The main contribution of this methods article is to introduce a visual data analytics framework, as summarized in Figure 1, to the IS literature, describing the steps from building a theoretical foundation, to constructing and validating features, and to conducting empirical studies. In doing so, this article makes the following specific contributions to the literature.

First, the proposed deep learning approach enables *large-scale* visual data analysis. This can have a significant impact not only on the IS and social media research but also on the fields of advertising, marketing, and psychology, which often examine the effect of visual content. Existent studies mainly

Table 11. Main Effects of Visual and Textual Features on Numbers of Reblogs and Likes

	(1)	(2)	(3)	(4)
	Negative Binomial		Linear	
Variables	Reblogs	*Likes*	*Log-Reblog*	*Log-Like*
ImageObjectComplexity	-0.0763**	-0.0558**	-0.0719***	-0.0492***
	(0.0366)	(0.0217)	(0.0198)	(0.0170)
ImagePixelComplexity	0.119***	0.0947**	0.0585**	0.0404
	(0.0411)	(0.0418)	(0.0290)	(0.0302)
AestheticScore	0.0641**	0.0380*	0.114***	0.0675***
	(0.0303)	(0.0205)	(0.0211)	(0.0183)
AdultContent	0.406**	0.409***	0.197***	0.305***
	(0.184)	(0.0967)	(0.0753)	(0.0889)
HasCelebrity	0.410***	0.366***	0.218***	0.254***
	(0.0989)	(0.0911)	(0.0625)	(0.0495)
TextTopicComplexity	0.000950	0.0150	-0.0160	-0.000899
	(0.0267)	(0.0189)	(0.0121)	(0.00946)
TextSentenceComplexity	-0.104*	-0.106***	-0.0352	-0.0257
	(0.0546)	(0.0368)	(0.0267)	(0.0222)
ImageConsistency	0.0566**	0.0644***	0.0591***	0.0613***
	(0.0226)	(0.0208)	(0.0174)	(0.0148)
TextConsistency	0.0845***	0.0644**	0.0718***	0.0500***
	(0.0270)	(0.0250)	(0.0188)	(0.0154)
ImageTextSimilarity	0.568**	0.329**	0.276***	0.150*
	(0.262)	(0.148)	(0.105)	(0.0826)
Other control variables	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
Blog fixed effects	Yes	Yes		
Constant	3.202***	4.492***	3.725***	3.140***
	(0.854)	(1.071)	(0.745)	(0.806)
Observations	34558	34558	34558	34558

Note: This table reports the baseline empirical results showing how different visual and textual features influence the consumer engagements in companies' social media posts. Columns (1)–(2) use negative binomial regression controlled for blog dummies. Columns (3)–(4) use fixed effects linear regression and the dependent variables are log-transformed numbers of reblogs and likes. Other control variables are omitted because the length of the page. Robust standard errors in parentheses: ***p < 0.01, **p < 0.05, *p < 0.1.

relied on hand-crafted visual features, which require significant manual engineering effort. Therefore, the empirical findings from relatively small-scale experiments, which involve limited numbers of handpicked picture samples, can be subject to generalizability issues. The proposed framework addresses this scalability issue.

Second, our approach enables us to construct *qualitatively new* visual content measures (e.g., object-level image complexity, image-text similarity, and generic image/text features) that could be difficult or impossible to operationalize with a traditional human coding approach. Our two case studies show that these new measures can play significant roles in predicting a social media post's popularity and in explaining the role of visual features on consumers' social media information processing.

Third, we provide practical guidance on *how* visual and textual content can be analyzed in social media research. Specifically, the paper introduces how open source ML frameworks, pre-trained deep learning models, and cloud services can be used to operationalize visual and textual content measures ("Visual and Textual Feature Construction" section and Table 2). The AMT study ("Validation of Visual and Textual Features" section) also shows how the measures constructed from deep learning and text mining can be validated with a crowdsourcing approach.

Finally, this paper can serve as an example of IS research that creates "synergies between Big Data and theory" (Rai 2016). There has been an active conversation in the IS community on the reconciliation of the relationship between data- and theory-driven research (Johnson et al. 2019; Maass et al.

2018). This article illustrates how the two can benefit from each other. The unstructured nature of big data creates a challenge for researchers in selecting and operationalizing the relevant constructs out of numerous possibilities. We draw on the ELM framework as a theoretical foundation in selecting visual measures that can affect social media ad effectiveness. We believe that, in turn, the findings from large-scale empirical analysis can facilitate further theoretical development.

Managerial Implications

The proposed framework also has practical implications to the managers who want to harness the value of big data. Whether in social media or traditional news media, the importance of visual content will continue to increase with technological advancements. Smart mobile devices are generating unprecedentedly large volumes of visual content, such as photos and videos, and faster Internet connectivity facilitates the active sharing of visual content among consumers. In addition, retail and fashion sectors are actively applying visual analytics to fashion data for personalized recommendations[20] and curation.[21] Moreover, emerging immersive technologies, such as virtual reality and augmented reality, will accelerate this trend.

In this visual content-oriented environment, managers are facing challenges making sense of visual data. We believe that the proposed analytics framework can help them "measure" these unstructured (textual and visual) data in a systematic and scalable manner. As exemplified by this paper, deep learning models can extract meaningful features (e.g., objects in images, aesthetic levels, adult-content scores) from large-scale content data with minimal manual human intervention. By automating the measurement process in social media management, firms can analyze huge volumes of online visual content and ads to make faster and more reliable data-driven decisions.

Societal Implications

ML and AI in general are increasingly playing major roles in our society and some of the automated tasks are critical ones, affecting our jobs, health, and legal systems. While the use of visual data analytics and ML can generally create significant value, any unintended negative societal impact, such as privacy, ethical, and accountability issues of deep learning and other ML approaches, should be carefully examined. For example, facial recognition, often enabled by deep learning techniques, can create privacy concerns for consumers.[22] Moreover, when biases in the training data are not properly accounted for, ML models can misuse protected characteristics (e.g., gender, age, ethnicity) and obstruct fairness, which recently created serious discrimination issues in critical situations such as sentencing in justice systems[23] and hiring in labor markets.[24] These real-life cases reveal the urgency to investigate accountability and transparency issues in deep learning models, which are often regarded as "blackbox" models (Abbasi et al. 2018; Diakopoulos 2016). However, we note that these crucial issues are less harmful in our setting and that we utilize deep learning models to generate human-interpretable features to better process massive amounts of visual content.

Conclusion and Future Directions

Visual content has grown to be an integral part of social media. Due to methodological challenges, studies in the social media literature have been constrained by simple or manually constructed features in analyzing unstructured data, especially visual content. In this paper, we take a step forward by proposing a deep learning-based visual data analytics framework to overcome this limitation. Specifically, we leverage a deep CNN that can algorithmically transform unstructured image data into useful representations for computer vision tasks. Combined with text mining techniques, this enabled us to construct novel features that would have previously been difficult to operationalize without manual coding. We evaluate the validity of the proposed visual and textual features and show their effectiveness with two case studies. This paper contributes to both academia and industry by extending researchers and companies' abilities to systematically investigate visual information, to understand consumer behavior, and potentially to enable informed decisions.

Because of the large scale of our study, we believe our framework and empirical results have broad applicability. Nevertheless, it is important to acknowledge limitations and possible future extensions. First, we think the proposed visual data analytics framework can be applied to other kinds of visual content (e.g., curating product and ad images, visual

[20]TechCrunch, "Amazon's new Echo Look Has a Built-in Camera for Style Selfies" (https://goo.gl/7HTrSL).

[21]Quartz, "Artificial Intelligence Can Say Yes to the Dress" (https://goo.gl/bnfvb4).

[22]*The New York Times*, "Facebook's Push for Facial Recognition Prompts Privacy Alarms" (https://goo.gl/URV3n6).

[23]Pro Publica, "Machine Bias" (https://goo.gl/KNBX4X).

[24]Reuters, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women" (https://goo.gl/hWuRZp).

inspection in manufacturing) besides social media images and text. In addition, the case studies only focus on a single type of visual data: images. Other visual data types such as videos can also be algorithmically analyzed with deep learning approaches (Hinton et al. 2012; Venugopalan et al. 2015). Furthermore, deep learning models and AI in general have started to demonstrate their capabilities in terms of high-level intelligence tasks that require sophisticated reasoning and intuition. Recent events with the game of Go have demonstrated the potential of deep learning models (Silver et al. 2017). We believe that, in the future, more complicated business decisions and strategies can benefit from AI and ML. This paper can serve as a stepping stone toward this direction (Jain et al. 2018).

Acknowledgments

We thank the senior editor, associate editor, and two reviewers for their guidance and constructive feedback during the review process. We appreciate for the helpful suggestions and comments from the audience at the 2015 WeB and WITS conferences and 2016 INFORMS, CIST, DSI, Texas FreshAIR, and UKC conferences as well as seminar participants at the Arizona State University, Chung-Ang University, Hanyang University, Korea University, Kyungpook National University, Kyung Hee University, Seoul National University, Simon Fraser University, Sungkyunkwan University, University of British Columbia, University of Connecticut, University of North Texas, University of Texas at Arlington, University of Utah, and Yonsei University. For helpful feedback, the authors greatly thank Wayne Hoyer and Sang Pil Han. We also thank Myunghwan Lee for his help in converting our LaTeX manuscript to Microsoft Word format.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," TensorFlow (https://www.tensorflow.org).

Abbasi, A., Li, J., Clifford, G., and Taylor, H. 2018. "Make 'Fairness by Design' Part of Machine Learning," *Harvard Business Review* (https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning).

Adomavicius, G., Bockstedt, J., and Curley, S. P. 2015. "Bundling Effects on Variety Seeking for Digital Information Goods," *Journal of Management Information Systems* (31:4), pp. 182-212.

Agrawal, J., and Kamakura, W. A. 1995. "The Economic Worth of Celebrity Endorsers: An Event Study Analysis," *Journal of Marketing* (59:3), pp. 56-62.

Arun, R., Suresh, V., Madhavan, C. V., and Murthy, M. N. 2010. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining: Advances in Knowledge Discovery and Data Mining*, pp. 391-402.

Attneave, F. 1954. "Some Informational Aspects of Visual Perception," *Psychological Review* (61:3), pp. 183-193.

Bandari, R., Asur, S., and Huberman, B. A. 2012. "The Pulse of News in Social Media: Forecasting Popularity," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, pp. 26-33.

Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. 2016. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction,* pp. 279-283.

Bengio, Y., Courville, A., and Vincent, P. 2013. "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (35:8), pp. 1798-1828.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Bloch, P. H. 1995. "Seeking the Ideal Form: Product Design and Consumer Response," *Journal of Marketing* (59:3), pp. 16-29.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. 2017. "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics* (5), pp. 135-146.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. 2009. "A Density-Based Method for Adaptive LDA Model Selection," *Neurocomputing* (72:7-9), pp. 1775-1781.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. 2014. "Return of the Devil in the Details: Delving Deep into Convolutional Nets," in *Proceedings of the British Machine Vision Conference,* Nottingham, UK.

Childers, T. L., and Houston, M. J. 1984. "Conditions for a Picture-Superiority Effect on Consumer Memory," *Journal of Consumer Research* (11:2), pp. 643-654.

Chollet, F. 2015. "Keras" (https://keras.io).

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. 2004. "Visual Categorization with Bags of Keypoints," in *Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1-22.

Deng, L., and Poole, M. S. 2010. "Affect in Web Interfaces: A Study of the Impacts of Web Page Visual Complexity and Order," *MIS Quarterly* (34:4), pp. 711-730.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 4171-4186.

Dhar, S., Ordonez, V., and Berg, T. L. 2011. "High Level Describable Attributes for Predicting Aesthetics and Interestingness," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1657-1664.

Diakopoulos, N. 2016. "Accountability in Algorithmic Decision Making," *Communications of the ACM* (59:2), pp. 56-62.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. 2014. "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *Proceedings of the 31st International Conference on Machine Learning,* pp. 647-655.

Donderi, D. C. 2006a. "An Information Theory Analysis of Visual Complexity and Dissimilarity," *Perception* (35:6), pp. 823-835.

Donderi, D. C. 2006b. "Visual Complexity: A Review," *Psychological Bulletin* (132:1), pp. 73-97.

Donderi, D. C., and McFadden, S. 2005. "Compressed File Length Predicts Search Time and Errors on Visual Displays," *Displays* (26:2), pp. 71-78.

Edell, J. A., and Staelin, R. 1983. "The Information Processing of Pictures in Print Advertisements," *Journal of Consumer Research* (10:1), pp. 45-61.

Fong, N. M. 2017. "How Targeting Affects Customer Search: A Field Experiment," *Management Science* (63:7), pp. 2353-2364.

Fontanini, G., Bertini, M., and Del Bimbo, A. 2016. "Web Video Popularity Prediction Using Sentiment and Content Visual Features," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval,* pp. 289-292.

Forsythe, A., Nadal, M., Sheehy, N., Conde, C. J. C., and Sawey, M. 2011. "Predicting Beauty: Fractal Dimension and Visual Complexity in Art," *British Journal of Psychology* (102:1), pp. 49-70.

Friedman, H. H., and Friedman, L. 1979. "Endorser Effectiveness by Product Type," *Journal of Advertising Research* (19:5), pp. 63-71.

Geissler, G. L., Zinkhan, G. M., and Watson, R. T. 2006. "The Influence of Home Page Complexity on Consumer Attention, Attitudes, and Purchase Intent," *Journal of Advertising* (35:2), pp. 69-80.

Gelli, F., Uricchio, T., Bertini, M., Bimbo, A. D., and Chang, S. 2015. "Image Popularity Prediction in Social Media Using Sentiment and Context Features," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference,* pp. 907-910.

Gong, J., Abhisek, V., and Li, B. 2018. "Examining the Impact of Keyword Ambiguity on Search Advertising Performance: A Topic Model Approach," *MIS Quarterly* (42:3), pp. 805- 829.

Goodman, K. S. 1967. "Reading: A Psycholinguistic Guessing Game," *Literacy Research and Instruction* (6:4), pp. 126-135.

Gough, P. B. 1972. "One Second of Reading," *Visible Language* (6:4), pp. 291-320.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. 2018. "Learning Word Vectors for 157 Languages," in *Proceedings of the 11th International Conference on Language Resources and Evaluation,* pp. 3483-3487.

Griffiths, T. L., and Steyvers, M. 2004. "Finding Scientific Topics," *Proceedings of the National Academy of Sciences* (101:Suppl. 1), pp. 5228-5235.

Gunning, R. 1952. *The Technique of Clear Writing*, New York: McGraw-Hill.

He, K., Zhang, X., Ren, S., and Sun, J. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE International Conference on Computer Vision,* pp. 1026-1034.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 770-778.

Heinzerling, B., and Strube, M. 2018. "BPEmb: Tokenization-Free Pre-trained Subword Embeddings in 275 Languages," in *Proceedings of the 11th International Conference on Language Resources and Evaluation,* pp. 2989-2993.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., and Sainath, T. N. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine* (29:6), pp. 82-97.

Huhmann, B. A. 2003. "Visual Complexity in Banner Ads: The Role of Color, Photography, and Animation," *Visual Communication Quarterly* (10:3), pp. 10-17.

Jain, H., Padmanabhan, B., Pavlou, P. A., and Santanam, R. T. 2018. "Call for Papers: Special Issue of Information Systems Research—Humans, Algorithms, and Augmented Intelligence: The Future of Work, Organizations, and Society," *Information Systems Research* (29:1), pp. 250-251.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. 2014. "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia,* pp. 675-678.

Jiang, Z., Wang, W., Tan, B. C. Y., and Yu, J. 2016. "The Determinants and Impacts of Aesthetics in Users' First Interaction with Websites," *Journal of Management Information Systems* (33:1), pp. 229-259.

Johnson, M. D., Herrmann, A., and Huber, F. 2006. "The Evolution of Loyalty Intentions," *Journal of Marketing* (70:2), pp. 122-132.

Johnson, S. L., Gray, P., and Sarker, S. 2019. "Revisiting IS Research Practice in the Era of Big Data," *Information and Organization* (29:1), pp. 41-56.

Kamins, M. A. 1989. "Celebrity and Noncelebrity Advertising in a Two-Sided Context," *Journal of Advertising Research* (29:3), pp. 34-42.

Kelman, H. C. 1961. "Processes of Opinion Change," *Public Opinion Quarterly* (25:1), pp. 57-78.

Kim, M., and Lennon, S. 2008. "The Effects of Visual and Verbal Information on Attitudes and Purchase Intentions in Internet Shopping," *Psychology & Marketing* (25:2), pp. 146-178.

Kincaid, J. P., Fishburne Jr., R. P., Rogers, R. L., and Chissom, B. S. 1975. "Derivation of New Readability Formulas (Automated Readability Index, FOG Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Technical Report, Institute for Simulation and Training, University of Central Florida.

Kintsch, W., and van Dijk, T. A. 1978. "Toward a Model of Text Comprehension and Production," *Psychological Review* (85:5), pp. 363-394.

Kosslyn, S. M. 1975. "Information Representation in Visual Images," *Cognitive Psychology* (7:3), pp. 341-370.

Kovashka, A., Russakovsky, O., Fei-Fei, L., and Grauman, K. 2016. "Crowdsourcing in Computer Vision," *Foundations and Trends in Computer Graphics and Vision* (10:3), pp. 177-243.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2017. "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM* (60:6), pp. 84-90.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. 2015. "Human-Level Concept Learning through Probabilistic Program Induction," *Science* (350:6266), pp. 1332-1338.

Lakkaraju, H., and Ajmera, J. 2011. "Attention Prediction on Social Media Brand Pages," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management,* pp. 2157-2160.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. 2016. "Neural Architectures for Named Entity Recognition," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 260-270.

Lane, V. R. 2000. "The Impact of Ad Repetition and Ad Content on Consumer Perceptions of Incongruent Extensions," *Journal of Marketing* (64:2), pp. 80-91.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436-444.

Lee, D., Hosanagar, K., and Nair, H. 2018. "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science* (64:11), pp. 5105-5131.

Lee, G. M., He, S., Lee, J., and Whinston, A. B. 2020. "Matching Mobile Applications for Cross-Promotion," *Information Systems Research* (31:3), pp. 865-891.

Lee, G. M., Qiu, L., and Whinston, A. B. 2016. "A Friend Like Me: Modeling Network Formation in a Location-Based Social Network," *Journal of Management Information Systems* (33:4), pp. 1008-1033.

Lerman, K., and Hogg, T. 2010. "Using a Model of Social Dynamics to Predict Popularity of News," in *Proceedings of the 19th International Conference on World Wide Web,* pp. 621-630.

Levi, G., and Hassner, T. 2015a. "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns," in *Proceedings of the ACM International Conference on Multimodal Interaction,* pp. 503-510.

Levi, G., and Hassner, T. 2015b. "Age and Gender Classification Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34-42.

Li, M., Wei, K.-K., Tayi, G. K., and Tan, C.-H. 2016. "The Moderating Role of Information Load on Online Product Presentation," *Information & Management* (53:4), pp. 467-480.

Lien, N.-H. 2001. "Elaboration Likelihood Model in Consumer Research: A Review," *Proceedings of the National Science Council Part C: Humanities and Social Sciences* (11:4), pp. 301-310.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. 2015. "Learning Transferable Features with Deep Adaptation Networks," in *Proceedings of the 32nd International Conference on Machine Learning,* pp. 97-105.

Lowe, D. G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision* (60:2), pp. 91-110.

Lv, J., Liu, W., Zhang, M., Gong, H., Wu, B., and Ma, H. 2017. "Multi-Feature Fusion for Predicting Social Media Popularity," in *Proceedings of the 25th ACM on Multimedia Conference,* pp. 1883-1888.

Ma, L., Sun, B., and Kekre, S. 2015. "The Squeaky Wheel Gets the Grease—An Empirical Analysis of Customer Voice and Firm Intervention on Twitter," *Marketing Science* (34:5), pp. 627- 645.

Maass, W., Parsons, J., Purao, S., Storey, V. C., and Woo, C. 2018. "Data-driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research," *Journal of the Association for Information Systems* (19:12), pp. 1253- 1273.

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., and Carballal, A. 2015. "Computerized Measures of Visual Complexity," *Acta Psychologica* (160), pp. 43-57.

MacInnis, D. J., and Price, L. L. 1987. "The Role of Imagery in Information Processing: Review and Extensions," *Journal of Consumer Research* (13:4), pp. 473-491.

Masi, I., Rawls, S., Medioni, G., and Natarajan, P. 2016. "Pose-Aware Face Recognition in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4838-4846.

Masi, I., Tran, A., Hassner, T., Leksut, J. T., and Medioni, G. 2016. "Do We Really Need to Collect Millions of Faces for Effective Face Recognition?," in *Proceediknqs of the European Conference on Computer Vision,* pp. 579-596.

McAlister, L. 1982. "A Dynamic Attribute Satiation Model of Variety-Seeking Behavior," *Journal of Consumer Research* (9:2), pp. 141-150.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. "Distributed Representations of Words and Phrases and their Compositionality," in *Proceedings of the Conference on Neural Information Processing Systems,* pp. 3111-3119.

Miniard, P. W., Bhatla, S., Lord, K. R., Dickson, P. R., and Unnava, H. R. 1991. "Picture-Based Persuasion Processes and the Moderating Role of Involvement," *Journal of Consumer Research* (18:1), pp. 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), pp. 12-24.

Mitra, B. 2015. "Exploring Session Context Using Distributed Representations of Queries and Reformulations," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 3-12.

Oliver, R. L. 1999. "Whence Consumer Loyalty?," *Journal of Marketing* (63:4, Suppl. 1), pp. 33-44.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. 2014. "Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1717-1724.

Page, C., and Herr, P. M. 2002. "An Investigation of the Processes by Which Product Design and Brand Strength Interact to Determine Initial Affect and Quality Judgments," *Journal of Consumer Psychology* (12:2), pp. 133-147.

Palmer, S. E. 1999. *Vision Science: Photons to Phenomenology*, Cambridge, MA: MIT Press.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. 2015. "Deep Face Recognition," in *Proceedings of the British Machine Vision Conference,* pp. 41.1-41.12.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. 2017. "Automatic Differentiation in PyTorch," in *Proceedings of the Conference on Neural Information Processing Systems Autodiff Workshop*.

Pennington, J., Socher, R., and Manning, C. D. 2014. "GloVe: Global Vectors for Word Representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* pp. 1532-1543.

Petty, R. E., and Cacioppo, J. T. 1986. "The Elaboration Likelihood Model of Persuasion," *Advances in Experimental Social Psychology* (19), pp. 123-205.

Petty, R. E., Cacioppo, J. T., and Schumann, D. 1983. "Central and Peripheral Routes to Advertising Effectiveness: The Moderating Role of Involvement," *Journal of Consumer Research* (10:2), pp. 135-146.

Phillips, B. J. 2000. "The Impact of Verbal Anchoring on Consumer Response to Image Ads," *Journal of Advertising* (29:1), pp. 15-24.

Pieters, R., Wedel, M., and Batra, R. 2010. "The Stopping Power of Advertising: Measures and Effects of Visual Complexity," *Journal of Marketing* (74:5), pp. 48-60.

Pieters, R., Wedel, M., and Zhang, J. 2007. "Optimal Feature Advertising Design Under Competitive Clutter," *Management Science* (53:11), pp. 1815-1828.

Rai, A. 2016. "Editor's Comments: Synergies Between Big Data and Theory," *MIS Quarterly* (40:2), pp. iii-ix.

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., Ng, A. Y., and Lungren, M. P. 2018. "Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists," *PLOS Medicine* (15:11), pp. 1-17.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., and Pettigrue, A. 2018. "Conversational AI: The Science Behind the Alexa Prize," in *Proceedings of the 1st Alexa Prize*.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. 2014. "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806-813.

Řehůřek, R., and Sojka, P. 2010. "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, pp. 45-50.

Ren, S., He, K., Girshick, R., and Sun, J. 2017. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (39:6), pp. 1137-1149.

Rosenholtz, R., Li, Y., and Nakano, L. 2007. "Measuring Visual Clutter," *Journal of Vision* (7:2), pp. 17-17.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. 2015. "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* (115:3), pp. 211-252.

Sengamedu, S. H., Sanyal, S., and Satish, S. 2011. "Detection of Pornographic Content in Internet Images," in *Proceedings of the 19th ACM International Conference on Multimedia,* pp. 1141-1144.

Severn, J., Belch, G. E., and Belch, M. A. 1990. "The Effects of Sexual and Non-sexual Advertising Appeals and Information Level on Cognitive Processing and Communication Effectiveness," *Journal of Advertising* (19:1), pp. 14-22.

Shi, Z., Lee, G. M., and Whinston, A. B. 2016. "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence," *MIS Quarterly* (40:4), pp. 1035-1056.

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. 2017. "Mastering the Game of Go Without Human Knowledge," *Nature* (550:7676), pp. 354-359.

Simonson, I. 1990. "The Effect of Purchase Quantity and Timing on Variety-Seeking Behavior," *Journal of Marketing Research* (27:2), pp. 150-162.

Simonyan, K., and Zisserman, A. 2015. "Very Deep Convolutional Networks for Large-scale Image Recognition," in *Proceedings of the International Conference on Learning Representations*.

Singh, P. V., Sahoo, N., and Mukhopadhyay, T. 2014. "How to Attract and Retain Readers in Enterprise Blogging?," *Information Systems Research* (25:1), pp. 35-52.

Smith, R. A. 1991. "The Effects of Visual and Verbal Advertising Information on Consumers' Inferences," *Journal of Advertising* (20:4), pp. 13-24.

Steadman, M. 1969. "How Sexy Illustrations Affect Brand Recall," *Journal of Advertising Research* (9:1), pp. 15-19.

Stieglitz, S., and Dang-Xuan, L. 2013. "Emotions and Information Diffusion in Social Media-Sentiment of Microblogs and Sharing Behavior," *Journal of Management Information Systems* (29:4), pp. 217-248.

Strebe, R. 2016. "Aesthetics on the Web: Effects on Approach and Avoidance Behaviour," *Behaviour & Information Technology* (35:1), pp. 4-20.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. 2016. "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2818-2826.

Taddy, M. 2015. "Document Classification by Inversion of Distributed Language Representations," in *Proceedings of the 53rd Annual Meeting of the Association of Computational Linguistics,* pp. 45-49.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. 2014. "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," in *Proceedings of the 52nd Annual*

Meeting of the Association for Computational Linguistics, pp. 1555-1565.

Teixeira, T., Wedel, M., and Pieters, R. 2012. "Emotion-Induced Engagement in Internet Video Advertisements," *Journal of Marketing Research* (49:2), pp. 144-159.

Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. 2009. "Visual Complexity of Websites: Effects on Users' Experience, Physiology, Performance, and Memory," *International Journal of Human–Computer Studies* (67:9), pp. 703-715.

Unnava, H. R., and Burnkrant, R. E. 1991. "An Imagery-Processing View of the Role of Pictures in Print Advertisements," *Journal of Marketing Research* (28:2), pp. 226-231.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. 2015. "Translating Videos to Natural Language Using Deep Recurrent Neural Networks," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 1494-1504.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. 2009. "Evaluation Methods for Topic Models," in *Proceedings of the 26th International Conference on Machine Learning,* pp. 1105-1112.

Wang, W., and Zhang, W. 2017. "Combining Multiple Features for Image Popularity Prediction in Social Media," in *Proceedings of the 25th ACM International Conference on Multimedia,* pp. 1901-1905.

Wu, B., Cheng, W., Zhang, Y., Huang, Q., Li, J., and Mei, T. 2017. "Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence,* pp. 3062-3068.

Wu, K., Vassileva, J., Zhao, Y., Noorian, Z., Waldner, W., and Adaji, I. 2016. "Complexity or Simplicity? Designing Product Pictures for Advertising in Online Marketplaces," *Journal of Retailing and Consumer Services* (28), pp. 17-27.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. 2014. "How Transferable Are Features in Deep Neural Networks?," in *Proceedings of the Conference on Neural Information Processing Systems,* pp. 3320-3328.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., and Snyder, M. 2016. "Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features," *Nature Communications* (7:12474).

Zadeh, A. H., and Sharda, R. 2014. "Modeling Brand Post Popularity Dynamics in Online Social Networks," *Decision Support Systems* (65), pp. 59-68.

Zhang, J., Liu, S., Li, M., Zhou, M., and Zong, C. 2014. "Bilingually-Constrained Phrase Embeddings for Machine Translation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* pp. 111-121.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. 2014. "Learning Deep Features for Scene Recognition Using Places Database," in *Proceedings of the Conference on Neural Information Processing Systems,* pp. 487-495.

About the Authors

Donghyuk Shin is an assistant professor in the Department of Information Systems at the W. P. Carey School of Business, Arizona State University. Prior to joining Arizona State University, he was a Machine Learning Scientist at Amazon Web Services. He received his Ph.D. in Computer Science from the University of Texas at Austin. His main research interests are in machine learning, deep learning, big data, recommender systems, social media, and business analytics applications. He has published in *MIS Quarterly* and top machine learning conferences including NeurIPS, ACM RecSys and CIKM.

Shu He is an assistant professor at the Department of Operations and Information Management, School of Business, University of Connecticut. She earned her Ph.D. in Economics from the University of Texas at Austin. Shu's research interests include social media, platform, online advertising, and cyber security. Her work has appeared in *Information Systems Research*, *MIS Quarterly*, and *Journal of Cybersecurity*. She has received grants from National Science Foundation and NET Institute to support her research.

Gene Moo Lee is an assistant professor of Information Systems at the Sauder School of Business, University of British Columbia. He received his Ph.D. in Computer Science from the University of Texas at Austin. His research interests are in the development and applications of Business Analytics. His papers have appeared in *MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Journal of Business Ethics*, and *Journal of Cybersecurity*. He has industry experiences at Samsung Electronics, AT&T, Intel, and Goldman Sachs. He holds eleven patents in mobile technology.

Andrew B. Whinston is the Hugh Cullen Chair Professor in the Information, Risk, and Operation Management Department at the McCombs School of Business at the University of Texas at Austin. He is also the director at the Center for Research in Electronic Commerce. He received his Ph.D. in Management from Carnegie Mellon University. His recent papers have appeared in *Information Systems Research*, *Journal of Management Information Systems*, *MIS Quarterly*, *Management Science*, *Marketing Science*, *Journal of Marketing*, and *Journal of Economic Theory*. He has published over 400 articles in refereed journals, 27 books, and 62 book chapters. In 2005, he received the Leo Award from the Association for Information Systems for his long-term research contribution to the information systems field. In 2009, he was named the Distinguished Fellow by the INFORMS Information Systems Society in recognition of his outstanding intellectual contributions to the information systems discipline. His Erdös number is 2.

Suleyman Cetintas is a Principal Research Scientist in the Advertising Science group at Yahoo Research. He received his Ph.D. in Computer Science from Purdue University, and his B.S. degree in Computer Engineering from Bilkent University. His research interests span a broad range of topics in online advertising, machine

learning, information retrieval, data and text mining, recommendation systems, and big data. He has published in *MIS Quarterly*, *Operations Research*, *Information Retrieval Journal*, *Journal of the Association for Information Science and Technology*, *IEEE Transactions on Learning Technologies*, among others.

Kuang-Chih Lee is the Head of Marketplace Optimization in AliExpress.com, an Alibaba Group company, the world's largest cross-border trading e-commerce platform with billions of dollars in transactions each year. He manages all aspects of research and development for real-time personalized e-commerce marketplace. His research interests span a broad range of topics in search and recommendation systems, online advertising, fraud detection, supply chain management, big data, distributed system, machine learning, data mining, NLP, and computer vision. As reported by Google Scholar, there are 5000+ citations to his publications. He received his CS Ph.D. from University of Illinois Urbana–Champaign in 2005.

Appendix A

List of Company Blogs by Industry Category

Automotive	acura, audicity, bmwusa, chopardclassicracing, departurelane, hondaloves, jeep, kia, landroverusa, lincolnmotorco, mercedesbenz, moversandmakers, sendthemasignal, smartownersbelike
Entertainment	aetv, beatsbydre, blackdiamondpa, conversemusic, disney, disneypixar, drmwrks, foxadhd, gamestop, gettyimages, hashtaglionsgate, hbo, hinl, huffingtonpost, hulu, ifc, latimes, listenforyourself, nbcnews, nbcnightlynews, newmuseum, npr, pbsdigitalstudios, pbstv, penguinteen, runningpress, sesamestreet, spotify, theatlantic, thedailyshow, theeconomist, ultimateears, vimeo, wmagazine, xbox, youtube
Fashion	10022-shoe, americanapparel, anthropologie, barbour, bergdorfgoodman, calvinklein, capitolcouture, cartier, clubmonaco, dior, dolcegabbana, donnasjournal, fancyfeast, glamour, goodarthlywd, gq, gucci, harpersbazaar, jcrew, katespadeny, lorealparisusa, maccosmetics, makeupforeverusa, maybelline, modcloth, olay, pfflyersstyle, ralphlauren, rayban, rickysnyc, sephora, stussy, suitsupply, teamtaylorswiftfragrances, timberland, topshop, urbanoutfitters, vanssnow, vogue, warbyparker
Finance	americanexpress, amexopenforum, bankrate, mastercard, yahoofinance
Food	americashamburgerhelper, amstellight, bemoretea, benandjerrys, coca-cola, cuttysark, dennys, digiorno, dqfanfood, earthsfinestguide, fruttarefruitbars, hellocereallovers, ihop, jr-watkins, kitkat, kraftrecipes, krispykreme, naturevalley, nowyourecooking, officialsubway, oreo, redbull, simplywonderful, skittles, smirnoffice, sprite, tacobell, tgifridays, usmacallan, wonkaicecream, wonkarandoms, zagat
Leisure	acehotel, adidasfootball, adidasoriginals, bandh, becausefutbol, enroutemagazine, holidayinn, lifeismagnifique, livelymorgue, lomographicsociety, lufthansa, montanamoment, nba, qatarairways, reebokclassics, starwoodhotels, takingoff, thescore, transformtomorrow, underarmour, visit-florida, whotels, yahoosports
Retail	archiemcphee, barbie, ebay, keds, macys, neimanmarcus, patagonia, sanborncanoecompany, thecorcorangroup10amspecial, theinsidesource, tiffanyandco, tjmaXXXX, vikingrange, yahooshopping
Technology	att, dell, generalelectric, gereports, ibmblr, ibmsocialbiz, madewithcode, marketr, mashablehq, norton, positivelytogether, smartercities, smarterplanet, sonos, sony, txchnologist, volition, yahoo, yahoolabs

Appendix B

Description on Convolutional Neural Networks

As deep learning is a relatively new machine learning approach to the information systems community, here we briefly introduce one of the deep learning approaches, convolutional neural network (CNN) (LeCun et al. 2015; Krizhevsky et al. 2017), which is widely used for image recognition and classification tasks in the computer vision literature.

The term *convolutional* originates from the convolution operator (or kernel) that preserves spatial relations between pixels and learns features using small local patches of the input image (e.g., 32-by-32 pixels). The main rationale behind the convolution operator is to exploit two important properties of visual data:

1. Locality: Objects in an image tend to have a local spatial support. That is, pixels in close proximity are more likely to be part of the same object compared to those that are far apart.

2. Translation Invariance: Object appearance is independent of location in the image. In other words, the same object can appear in different parts of an image.

Typically, a CNN model consists of multiple layers of neural network: input layer, hidden layers (including convolution layers), and output layer. The input layer reads the focal image to analyze and the output layer predicts the object categories that exist in the image. Each hidden layer transforms the representation from the previous layer into a more abstract representation, where the convolution operator is applied to the entire input image as a sliding window at the convolution layers. The key aspect of CNN is that it *automatically* discovers robust representations needed for accurate classification via the composition of such multiple transformations. In other words, the layers are not designed by humans (which is the case of most traditional image recognition and classification methods), but are learned from the data. The accuracy on a benchmark dataset is boosted to 97% with CNNs, whereas conventional methods with handcrafted features only achieved 72% accuracy.[25]

The CNN model we employ is a model developed at Yahoo! that drives many of its services including Flickr (photo service owned by the company). The architecture of the model is designed based on Donahue et al. (2014), Jia et al. (2014), Simonyan and Zisserman (2015), and Krizhevsky et al. (2017), which is trained using a proprietary Flickr dataset of more than 1.5 million images with 1,700 object categories. That is, given an image, the prediction is a 1,700-dimensional vector of confidence scores between 0 and 1 corresponding to each object category. The object categories are general enough to cover various types of objects and concepts (e.g., animals, people, electronics, food, furnishing, nature, vehicles, etc.) and balanced in terms of the number of images, similar to the benchmark ImageNet dataset (Russakovsky et al. 2015). This is particularly important as the deployed model should be able to handle a wide variety of images generated by users.

The second-to-last layer that feeds in to the final classification layer of the model has 4,096 nodes, whose activation outputs correspond to a 4,096-dimensional feature vector referred as *CNN codes*. These generic CNN code features are viewed as the final image representation learned by the model and have been shown to achieve superior generalization performance on various computer vision tasks (Donahue et al. 2014, Razavian et al. 2014, Yosinski et al. 2014, LeCun et al. 2015, He et al. 2015, Yu et al. 2016, Rajpurkar et al. 2018). We utilize the CNN codes in our predictive analytics case study.

As being utilized in Yahoo! services, the CNN model has been extensively tested to achieve generalization performance (i.e., accuracy on new test images) that meets the high standards of production-level services.[26] In fact, when we evaluated the CNN model performance on our Tumblr dataset, we found that the prediction accuracy is 91.9% (see Appendix E for details). While many new algorithms and architectures for CNN have been developed in the past couple of years (using more layers and different layer structure), such as InceptionNet (Szegedy et al. 2016) or ResNet (He et al. 2016), the overarching concept of CNN as described above remains largely unchanged.

[25]ImageNet Large Scale Visual Recognition Challenge (http://image-net.org/challenges/LSVRC).

[26]The specific performance of the model is part of Yahoo!'s intellectual property.

Appendix C

Description on LDA Topic Modeling

Latent Dirichlet Allocation (LDA) topic modeling (Blei et al. 2003) is an unsupervised generative probabilistic machine learning approach to discover latent topics from a collection of documents. Recently, it has been widely adopted in the information systems literature to analyze unstructured text data (Lee et al. 2016; Lee et al. 2020; Shi et al. 2016; Singh et al. 2014). The main idea of LDA is that documents are represented as random mixtures over a small number of (latent) topics, where a topic is characterized by a distribution over words. That is, documents are the realization of its underlying topics and LDA outputs the topic distribution for each document (i.e., post in our blog data). An important advantage of LDA is that researchers can read the automatically constructed keyword sets for each topic to understand the underlying topics. Operationally, we collect all text data from Tumblr posts and compute topic models using the LDA implementation in the GENSIM package (Řehůřek and Sojka 2010). The number of topics is an important hyperparameter of LDA, and we used multiple approaches suggested in the literature to find the appropriate number as shown in the following subsection.

Choosing Number of Topics in LDA

To determine an appropriate number of topics of LDA model in a systematic manner, we utilize multiple approaches suggested in the machine learning literature. In particular, we use (1) perplexity that quantifies how well the held-out test data are represented by the learned distributions (Wallach et al. 2009), (2) the estimated log-likelihood of data for a given number of topics (Griffiths and Steyvers 2004), (3) topic density based on distance between topics (Cao et al. 2009), and (4) the KL-divergence of salient distributions derived from a matrix factorization formulation of LDA (Arun et al. 2010). From Figure C1(a), which shows each of these criteria with varying number of topics, we can see that number of topics in the range of 20 to 40 are good candidates. In this study, we use 20-topic LDA model and validate the proposed topic-level text complexity measure in the "Validation of Visual and Textual Features" section against human coders. To further ensure the quality of our LDA model, we confirm that the resulting keyword sets form intuitively coherent topics.[27] Then, drawing on the existing works using topic-based text similarity (Lee et al. 2016, Shi et al. 2016), we construct each blog post's text consistency measure as well as image-text similarity measure.

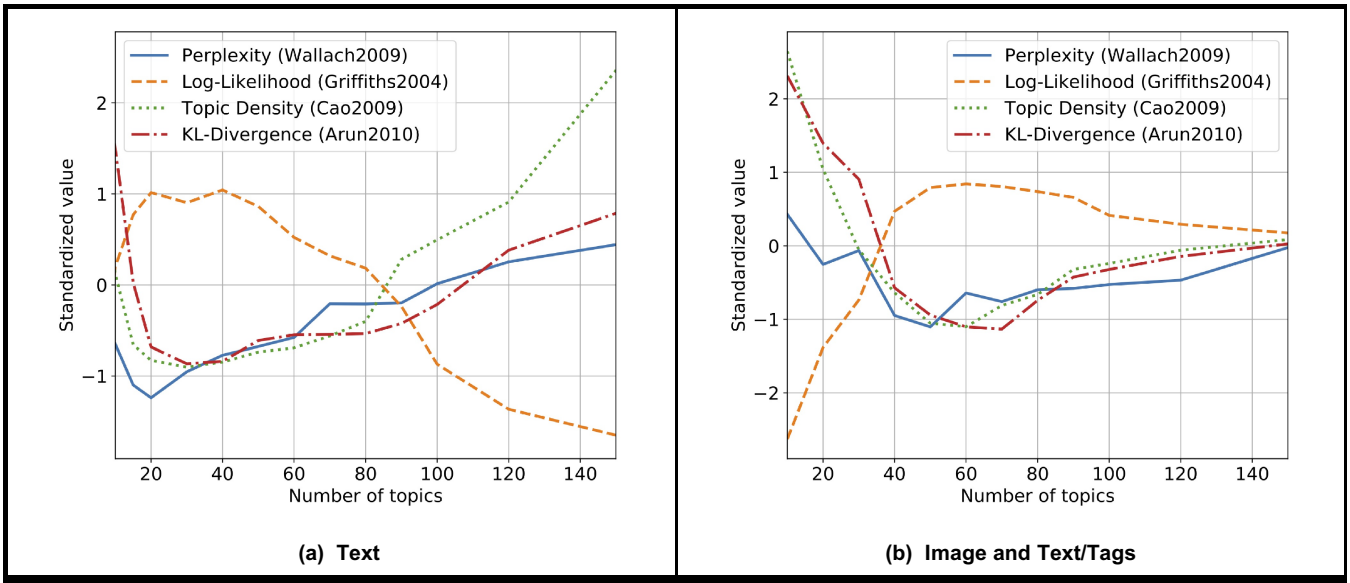


Figure C1. Different Criteria for Selecting the Number of Topics in LDA, Including Perplexity (Wallach et al. 2009), Topic Density (Cao et al. 2009), KL-Divergence (Arun et al. 2010) (Lower is Better), and Log-Likelihood (Griffiths and Steyvers 2004) (Higher is Better) for Text and Images with Text and Tags. Values are Standardized

[27] Topics and keywords constructed from text are available at https://goo.gl/Va7uTh.

Appendix D

Description on Word2vec Word Embedding

We use a word embedding approach to construct a micro-level text complexity measure, which to quantify the level of unpredictability of sentences of a post. Recently, a neural network inspired model called *word2vec* has been proposed that embeds words in a latent factor space such that it captures a large number of precise syntactic and semantic word relations (Mikolov et al. 2013). Learning such a distributed representation of words in a vector space has been successfully used in various natural language processing tasks (Lample et al. 2016; Mitra 2015; Taddy 2015; Tang et al. 2014; Zhang et al. 2014). Specifically, word2vec utilizes the technique called skip-gram with negative samples, which tries to represent each of the words by a *d*-dimensional vector so that words used in many similar contexts are close to each other in the vector space. This representation is accomplished by maximizing the predicted probability of words co-occurring within a small window of consecutive words in the training corpus (e.g., five words before and after the focal word). Mathematically, the word2vec model maximizes the following objective for all sentences is

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j\neq i,\, j=i-b}^{i+b}\log p\left(s_j \mid s_i\right) \tag{3}$$

where T is the number of words in sentence s, b is the window size, s_i is the vector representation of the i^{th} word in sentence s, and $p(s_j|s_i)$ is a neural network model.[28]

In contrast with LDA, which captures document-level associations, word2vec focuses on *local* context information. That is, word2vec predicts a nearby word given a particular word (focal word → nearby words), whereas LDA *globally* predicts words at the document level (document → topics → words). Another important difference is that the order of words has a significant impact in word2vec, whereas LDA uses a document/word-frequency matrix representation (i.e., bag-of-words) that ignores such ordering.

Appendix E

CNN Model Validation Using Amazon Mechanical Turk

We evaluated the performance of the CNN model used in this study on Tumblr images with human coders from AMT. The goal is to evaluate whether the labels predicted by the CNN model correctly match the image contents, which is a much simpler task than annotating unlabeled images that requires expert domain knowledge from human coders or various verification steps (Kovashka et al. 2016). Following the standard top-5 prediction accuracy metric of the ImageNet image recognition challenge (Russakovsky et al. 2015), we gave human coders an image and its top-5 predicted labels obtained by the CNN model. Then we asked how many of the labels describe the contents or some relevant context of the image or the image itself. When at least one of the AMT workers answered zero labels (i.e., none of the labels match the image), we consider it as an *incorrect* prediction. The other case (i.e., all human coders answered a positive number of labels match the image) is counted as a *correct* prediction. About 230 distinct workers participated in our AMT experiment, where each image was assigned to at least 5 workers. We employed the same best practices for AMT experiments as discussed in Appendix F to ensure the quality of the results.

Using a stratified random sample of 2,500 Tumblr images, we found the prediction accuracy to be 91.9%, where the mean and median of the number of labels that match the image are 3.27 and 3, respectively. In addition, if we further expand incorrect predictions to also include cases where at least one of the AMT workers answered only one of the labels matches the image, the accuracy slightly decreases to 86.2%. The results show that the CNN model used in this study predicts image labels with a very high accuracy, which aligns with accuracies reported in the ImageNet challenge on benchmark datasets (Russakovsky et al. 2015). One of the contributing factors of such good performance is the fact that approximately 24% of images from our Tumblr dataset are actually hosted on Flickr, which is the dataset used to train the CNN model. The results also reinforce existing studies showing the superior generalization performance of CNNs on different datasets or domains (Chatfield et al. 2014, Donahue et al. 2014, Razavian et al. 2014, Yosinski et al. 2014).

[28] We refer the reader to Mikolov et al. (2013) for details of the model.

Appendix F

Visual and Textual Feature Validation Using Amazon Mechanical Turk

Below we describe the details of our AMT survey instrument. For each human coder, we showed a pair of images or bodies of text (or both) from two different posts and asked the following questions for each of the targeted content measures:

- Object-Level Image Complexity: Pick the image that feels simpler or uniform.
- Image Aesthetic Score: Pick the image that has higher quality/aesthetic value or is more appealing.
- Image Text Similarity: Pick the post (image with text/tags) that feels more unified or coherent.
- Topic-Level Text Complexity: Pick the text that talks about more things or topics.
- Sentence-Level Text Complexity: Pick the text that is harder to read or feels unnatural.

To ensure the quality of workers and survey results, we followed some of the known best practices for AMT experiments identified in the literature (Lee et al. 2018):

- Each pair was assigned to at least five different workers. The option that received the majority vote from the workers is considered as the final selected option.
- We restrict our survey to workers who at least completed 50 tasks and had a 97% or better task-approval rate.
- We use only workers from the U.S. to filter out those potentially not proficient in English, and to closely match the geographic of our data (recall that majority of the blogs are targeting U.S. consumers).
- We refined our survey instrument through an iterative series of trial runs on small batches of the pairs.
- Our survey instrument includes only simple questions and most of the text are relatively short. On average, we found that it took about 7 seconds for each task. We defined less than 5 seconds to be too short and discarded any pairs with completions times shorter than this.