



## Marketing Science

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Capturing Changes in Social Media Content: A Multiple Latent Change-point Topic Model

Ning Zhong, David A. Schweidel

To cite this article:

Ning Zhong, David A. Schweidel (2020) Capturing Changes in Social Media Content: A Multiple Latent Change-point Topic Model. Marketing Science 39(4):827-846. <https://doi.org/10.1287/mksc.2019.1212>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2020, INFORMS

Please scroll down for article—it is on subsequent pages





With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Capturing Changes in Social Media Content: A Multiple Latent Changepoint Topic Model

Ning Zhong,<sup>a</sup> David A. Schweidel<sup>b</sup>

<sup>a</sup>Smeal College of Business, Pennsylvania State University, University Park, Pennsylvania 16802; <sup>b</sup>Goizueta Business School, Emory University, Atlanta, Georgia 30322

Contact: [nuz85@psu.edu](mailto:nuz85@psu.edu),  <https://orcid.org/0000-0002-2417-0448> (NZ); [dschweidel@emory.edu](mailto:dschweidel@emory.edu),  <https://orcid.org/0000-0003-2665-3272> (DAS)

Received: October 14, 2016

Revised: January 20, 2018; April 9, 2019;  
July 17, 2019

Accepted: September 16, 2019

Published Online in Articles in Advance:  
March 9, 2020

<https://doi.org/10.1287/mksc.2019.1212>

Copyright: © 2020 INFORMS

**Abstract.** Although social media has emerged as a popular source of insights for both researchers and practitioners, much of the work on the dynamics in social media has focused on common metrics such as volume and sentiment. In this research, we develop a changepoint model to capture the underlying shifts in social media content. We extend latent Dirichlet allocation (LDA), a topic modeling approach, by incorporating multiple latent changepoints through a Dirichlet process hidden Markov model that allows for the prevalence of topics to differ before and after each changepoint without requiring prior knowledge about the number of changepoints. We demonstrate our modeling framework using social media posts from brand crises (Volkswagen's 2015 emissions testing scandal and Under Armour's 2018 data breach) and a new product launch (Burger King's 2016 launch of the Angriest Whopper). We show that our model identifies shifts in the conversation surrounding each of these events and outperforms both static and other dynamic topic models. We demonstrate how the model may be used by marketers to actively monitor conversations around their brands, including distinguishing between changes in the conversation arising from a shift in the contributor base and underlying changes in the topics discussed by contributors.

**History:** K. Sudhir served as the senior editor and Olivier Toubia served as associate editor for this article.

**Supplemental Material:** Data and the e-companion are available at <https://doi.org/10.1287/mksc.2019.1212>.

**Keywords:** social media • changepoint models • text analysis • topic models

## Introduction

Firms are increasingly interested in extracting marketing insights from social media data. The CMO Survey reported that more than 40% of surveyed companies are planning to invest in social listening and social media analytics.<sup>1</sup> According to Forrester Research's evaluation of leading social listening platforms, the top uses for social listening platforms included monitoring the brand and brand health, measuring campaign success, and better understanding customers (Ngo and Pilecki 2016).

Although much research on social media has focused on quantifiable measures (i.e., structured data) such as volume and sentiment, these metrics offer limited insights into how customers perceive the brand. For brands that monitor social media conversations in which the brand and its competitors are mentioned, volume and sentiment provide a convenient summary of how social media activity varies from one day to the next. Such measures have been the focus of research investigating the dynamics in online conversations (e.g., Godes and Silva 2012, Moe and Schweidel 2012, Schweidel and Moe 2014, Xiong and

Bharadwaj 2014, Ma et al. 2015, Borah and Tellis 2016, Fossen and Schweidel 2017), with limited work considering dynamics in the textual content of social media posts.

In this research, we contribute to the growing research on social media dynamics by developing a modeling framework to examine how the content of social media contributions may shift over time. As the foundation of our model, we employ latent Dirichlet allocation (LDA; Blei et al. 2003), a popular topic modeling framework that assumes documents are comprised of latent topics and that words are associated with different topics (e.g., Tirunillai and Tellis 2014, Büschken and Allenby 2016, Puranam et al. 2017). We combine the topic modeling framework with a hidden Markov model (HMM; e.g., Chib 1998, Netzer et al. 2008, Schweidel et al. 2011) that assumes the content of posts occurring before and after each changepoint may differ without requiring prior knowledge about the number of changepoints.

To the best of our knowledge, this research is the first to identify the presence and timing of latent changepoints in topics discussed in social media messages.

In doing so, our research complements prior work that examines differences in online conversations before and after specific known events such as the enactment of a new public policy (e.g., Puranam et al. 2017), and reveals how conversations evolve afterward. We demonstrate how our modeling approach can provide managers with an indication of when underlying shifts in social media conversations have occurred and investigate whether they stem from the participation of new contributors or changes in the topics discussed by existing contributors. By distinguishing between social media posts contributed by consumers who have previously engaged with the brand and those who are just joining the conversation, marketers can assess if the observed changes in the social media conversations are arising from differences in the contributor base or changes in brand perceptions, which may in turn affect how marketers choose to respond.

We demonstrate our modeling framework by applying it to three empirical contexts. We first investigate social media conversations mentioning Volkswagen around the time of its emissions testing scandal (Woodyard 2015). Although the timing of the event is known after the fact, our model is able to identify the changes in the conversation as they occur. Our analysis not only reveals that topics related to this brand crisis became more prevalent in social media posts mentioning the brand, but also sheds light on those topics that were less frequently mentioned following the changepoint. Having demonstrated the ability of the model to detect changepoints in the conversation, we then apply our model to two events that did not garner the same degree of media attention: Under Armour's 2018 data breach and Burger King's launch of the Angriest Whopper in 2016. Using Twitter data for these two contexts, we demonstrate how brands can employ our framework to assess if the observed conversation dynamics are due to shifts in the prevalence of topics among existing contributors or changes to the contributor base.

The contribution of this research lies in extending the use of topic models in marketing to capture changes in conversations that may occur over time. We develop a topic model with multiple latent changepoints to identify when and to what extent the shifts in the underlying topics mentioned occur. In doing so, we contribute to the literature that has investigated social media dynamics but primarily focused on metrics such as volume and sentiment (e.g., Godes and Silva 2012, Moe and Schweidel 2012, Borah and Tellis 2016). Our research also contributes to the growing body of research in marketing that seeks to leverage unstructured textual data for the purposes of marketing insights (e.g., Lee and Bradlow 2011, Netzer et al. 2012, Tirunillai and Tellis 2014) by incorporating

latent regimes into the data generating process. We illustrate the managerial relevance of the modeling framework by using it to detect conversational shifts. Moreover, we demonstrate how our approach can be used to diagnose if these shifts arise from changes in the contributor base or conversational shifts among the same contributors, which may affect how marketers choose to react.

The remainder of this manuscript proceeds as follows. We next review the related literature. We then describe our modeling framework and the metrics of interest to managers. We proceed to discuss the data sets employed in our empirical analysis. We then present our empirical findings and conclude with a discussion of potential areas for future research.

## Related Literature

We draw on multiple streams of work that have largely evolved independently of each other. We first discuss related work into the dynamics of consumers' social media activity. We then discuss research in marketing that has delved into user-generated content with a focus on text analytics methods. We then discuss the limited research that has examined the textual dynamics and the marketing literature on which we draw to develop a discrete state model of social media content.

### Social Media Dynamics

Although prior research has examined temporal patterns in social media activity, much of this stream has focused on metrics such as volume and sentiment. For example, Godes and Silva (2012) use product-level data to investigate the temporal and sequential evolution of online product ratings. Using individual-level data on online product reviews, Moe and Schweidel (2012) model a user's decision of whether to contribute a review, as well as the sentiment of the review. The authors demonstrate dynamics in users' incidence and evaluation decisions arising from heterogeneity across users. Schweidel and Moe (2014) also document the presence of dynamics in the sentiment expressed and the venue to which social media posts are contributed.

Research has also viewed product reviews as means by which early purchasers may provide potential buyers with more information than was initially available. Kuksov and Xie (2010) examine the impact that product reviews may have on the firm's pricing decisions. Sun (2012) looks at the impact of a high variance in previously contributed reviews as providing information to consumers when the average ratings is low, as this may indicate that the product appeals to some customers but not all. Moe and Trusov (2011) also examine how previously contributed reviews affect sales. In doing so, they decompose the effects of previous reviews into a direct effect on sales and

an indirect effect through their impact on subsequent reviews.

Understanding the dynamics present in social media activity is essential to maintaining the brand, sensing market, and managing customer relationships. Schweidel and Moe (2014) demonstrate that the analysis of social media data can yield a measure of brand health that is a leading indicator of survey-based metrics. Looking at how brand perceptions within an entire industry may shift, Borah and Tellis (2016) investigate the dynamics surrounding social media conversations following product recalls and find evidence of negative spillover effects. To spot market trends, Du and Kamakura (2012) propose a dynamic factor-analytic model that teases out the cross-brands common trend underlying Google trends data. In terms of managing relationships, Ma et al. (2015) use an HMM to characterize customers' relationship with the firm and their social media activity that mentions the firm. The authors show how their model can inform the firm's decision of for which customers it will intervene to improve the relationship.

As our discussion of the extant research on social media dynamics reveals, there is considerable interest in identifying social media dynamics so that brands may take appropriate actions. Despite some exceptions (e.g., Liu et al. 2016, 2017), much of the extant research has focused on metrics such as volume, sentiment, and variation, which has not sought to examine the dynamics in the content of social media posts.

### Text Analysis of User-Generated Content

Text analysis has become more popular in the marketing literature in recent years. Researchers have applied text analytic methods to data arising from social tags (Nam and Kannan 2014, Nam et al. 2017), microblogs (Culotta and Cutler 2016), forums (Netzer et al. 2012), online reviews (Tirunillai and Tellis 2014, Büschken and Allenby 2016), and search data (Ringel and Skiera 2016, Liu and Toubia 2018). To extract meaningful content from text, some work takes advantage of word-level or phrase-level text to cluster and summarize topics by similarity or co-occurrence. For example, Archak et al. (2011) incorporate the phrases of product reviews into a consumer choice model. Lee and Bradlow (2011) develop an automated approach to use the text of product reviews to conduct marketing research. Their approach begins by identifying specific phrases that are present in an online review and then grouping together keywords that are similar to each other. In a similar fashion, Netzer et al. (2012) look at the frequency with which brands are comentioned, a high tendency for brands to be

comentioned as indicative of the brands being competitive in the minds of consumers.

More recently, marketing researchers have begun to use and extend LDA (Blei et al. 2003), a topic modeling framework. LDA assumes that for each token position within a document, a latent topic is drawn. Conditional on the topic drawn, a word is then drawn from a vocabulary. Tirunillai and Tellis (2014) build upon the LDA framework by simultaneously identifying the dimensions of brands that are discussed in user-generated product reviews and the sentiment associated with the dimension. Büschken and Allenby (2016) extend LDA by developing a sentence-constrained LDA that they show provides a superior model fit compared with the standard LDA model in prediction. To investigate the consumers' content preference, Liu and Toubia (2018) develop a hierarchical dual LDA that relates the topics in search queries to the topics in webpages of top search results. Other researchers have used LDA to investigate how the content of social media messages has changed over time, specifically in regard to known events. For example, Puranam et al. (2017) use LDA to investigate the effect of new regulations on the content of online reviews.

Though temporal changes have been examined using LDA-based models, such approaches are limited in their ability to detect emerging shifts in the content of social media messages and alert managers. Although Tirunillai and Tellis (2014) examine the frequency with which brands are associated with specific dimensions and how this frequency shifts over time, they assume a time-invariant process governing the content of reviews. That is, they assume that the parameters governing the prevalence of topics do not change over time. Similarly, Puranam et al. (2017) assume a time-invariant data generating process and conduct a difference-in-difference analysis to estimate the impact of the regulations. Though the authors examine how topic prevalence differs before and after an event, such an approach can only be applied after the fact when the event is known. This limits the ability of managers to react to shifts in the topics being mentioned.

To the best of our knowledge, there has been limited research that incorporates dynamics into the underlying process that governs textual content. Blei and Lafferty (2006) propose a dynamic topic model in which the hyperpriors of topic proportions and word distributions follow random walks. Puranam et al. (2017) test this model and find little evidence of topic evolution in their context. In the context of customer relationship management, researchers have discussed the flexibility afforded by capturing state



dependence through discrete states rather than through a continuous specification (e.g., Netzer et al. 2008). Whereas allowing for the continuous evolution of the hyperpriors (e.g., Blei and Lafferty 2006) and the use of latent changepoints both allow one to incorporate dynamics into an LDA-based model, we favor the use of latent changepoints for the ease with which it can be implemented using Gibbs sampling (e.g., Griffiths and Steyvers 2004) and the interpretability that it affords. In particular, our approach facilitates comparing the prevalence of topics across all the regimes. As we demonstrate, this provides managers with a convenient means of assessing when and to what extent the conversation has shifted.

## Model Development

In this section, we describe the LDA-based modeling framework that we develop to detect shifts in the content of social media posts. We begin by briefly introducing the standard LDA model (e.g., Blei et al. 2003, Griffiths and Steyvers 2004) that is at the core of our model. Building upon this foundation, we then embed a multiple latent changepoint model to accommodate the temporal nature by which social media posts are made. Our resulting LDA with multiple latent changepoints (LDA-MLC) model allows for the content of social media posts to manifest from multiple underlying regimes.

### Latent Dirichlet Allocation

LDA is a generative statistical model in the field of topic modeling that portrays collections of discrete data, such as a text corpus, by a finite mixture of unobserved clusters, such as underlying topics (Blei et al. 2003). It allows each document in a text collection to be described by a mixture of topics, with each word being attributed to a topic with different weight. We refer to literature on natural language processing and define the following terms: a *word* is a set of English characters in a vocabulary generated from a document collection. A *token* is a chopped instance of a sequence of characters in documents, which in our case can be a word or an acronym. A *document*, or social media post in our empirical analysis, is a sequence of tokens that may consist of repeated words. To provide an illustration, consider the following post about the Volkswagen emissions cheating scandal:

Volkswagen has said that in some cases, the cars can be fixed by reprogramming the software. But in other cases, Volkswagen may need to install new hardware. Matthias Müller, the Volkswagen chief executive, has not ruled out giving some customers new vehicles if repairs are not possible. Hang on to your tdi's... this could be the biggest rfd deal ever!!! (hipster 2015)

The document is comprised of four sentences. The word “Volkswagen” appears three times in the document: the first token of the first sentence, the fifth token of the second sentence, and the fourth token of the third sentence.

For each document  $d$  in the document collection, LDA assumes the following generative process such that:

1. Choose  $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$ ; and
2. For each token  $i$  in document  $d$ ,
  - (a) choose a topic  $z_i \sim \text{Multinomial}(\theta^{(d)})$ , and
  - (b) choose a word  $w_i | z_i \sim \text{Multinomial}(\phi^{(z_i)})$ ,

where  $\theta^{(d)}$  is a document-specific probability vector over topics that denotes topic proportions in document  $d$ . The topic-specific probability vector  $\phi^{(z)}$  for topic  $z$  consists of the probabilities with which words are drawn for a token associated with topic  $z$ . Griffiths and Steyvers (2004) propose the use of Gibbs sampling for LDA by incorporating a Dirichlet prior on  $\phi^{(j)}$  such that  $\phi^{(j)} \sim \text{Dirichlet}(\beta)$ . Rather than explicitly estimating  $\theta$  and  $\phi$ , they evaluate the posterior distribution over the assignment of words  $w$  to topics  $z$ . As detailed by Griffiths and Steyvers (2004), the joint distribution  $P(w, z) = P(w|z)P(z)$  can be computed by integrating out  $\theta$  and  $\phi$ , where

$$P(w|z) = \left( \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \right)^J \prod_{j=1}^J \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)}, \quad (1)$$

$$P(z) = \left( \frac{\Gamma(J\alpha)}{\Gamma(\alpha)^J} \right)^D \prod_{d=1}^D \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n^{(d)} + J\alpha)}, \quad (2)$$

in which  $n_j^{(w)}$  counts the number of times word  $w$  is assigned to topic  $j$  in the whole collection,  $n_j^{(\cdot)}$  counts the number of tokens assigned to topic  $j$  in the whole collection,  $W$  is the size of vocabulary,  $n_j^{(d)}$  counts the number of tokens assigned to topic  $j$  in document  $d$ ,  $n^{(d)}$  counts the number of all tokens in document  $d$ ,  $J$  is the number of topics,  $D$  is the number of documents, and  $\Gamma(\cdot)$  is a gamma function. The conditional distribution of token  $i$  being assigned to topic  $j$  can be represented as (Griffiths and Steyvers 2004):

$$P(z_i = j | z_{-i}, w, \alpha, \beta) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} (n_{-i,j}^{(d_i)} + \alpha), \quad (3)$$

where  $d_i$  and  $w_i$  denote the document and word index of token  $i$  respectively;  $n_{-i,j}^{(w_i)}$  counts the number of times word  $w$  is assigned to topic  $j$ , excluding the current token  $i$ , and  $n_{-i,j}^{(\cdot)}$  counts the number of times any token is assigned to topic  $j$ , excluding the current token  $i$ ;  $n_{-i,j}^{(d_i)}$  counts the number of times topic  $j$  is assigned to all tokens in document  $d$ , excluding the current token  $i$ , and  $n_{-i,j}^{(d)}$  counts the number of all tokens in document  $d$ , excluding the current

token  $i$ . The estimates of word distribution by topic and topic distribution by document can be obtained by

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}, \quad (4)$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n^{(d)} + J\alpha}, \quad (5)$$

where  $n_j^{(w)}$  counts the number of times word  $w$  is assigned to topic  $j$ , and  $n_j^{(\cdot)}$  counts the number of times any token is assigned to topic  $j$ ;  $n_j^{(d)}$  counts the number of times topic  $j$  is assigned to any tokens in document  $d$ , and  $n^{(d)}$  counts the number of all tokens in document  $d$ .

### Latent Dirichlet Allocation with Multiple Latent Changepoints

LDA assumes that the topic proportions of all documents in a text collection are drawn from a single underlying Dirichlet distribution, thereby ignoring the temporal nature of social media activity. Either exogenous shocks, such as brand crises, new product introductions and public policy changes, or endogenous evolution, such as user-generated conversations, may change the frequency with which topics are discussed in social media posts. If such events are known a priori, we may simply divide the documents into two sets: those documents that were contributed before the event and those documents that were contributed afterward (e.g., Puranam et al. 2017). If the event is unknown to us a priori, we cannot divide the data into two regimes. Moreover, the topic proportions may continue to change in the aftermath of such an event, requiring more than before and after regimes to capture temporal variation in conversations.

To overcome this limitation and detect changepoints as they emerge, we assume that the frequency with which topics are discussed may change over time. We allow for one or more underlying regimes from which the prevalence with which topics are discussed in a document, indicated by the vector of Dirichlet parameter  $\alpha$ , is drawn, while fixing the word distribution of each topic within the observation time window. We assume that the prevalence with which topics are discussed between each latent changepoint may differ. In our LDA-MLC model, we assume that there are multiple regimes that govern the content of social media posts, and that the regime at time  $t$  (denoted  $s_t$ ) evolves according to an HMM (e.g., Chib 1998, Netzer et al. 2008). Specifically, we assume the following generative process:

1. Choose a regime  $k \sim P(s_t | s_{t-1})$ , where  $s_t$  is the discrete state at time  $t$ ,  $t = 1, 2, \dots, T$ . Documents contributed at time  $t$  belong to regime  $k$ ;

2. For each document  $d$  in regime  $k$ , the collection of which is denoted  $\mathcal{C}_k$  with  $k \in \mathbb{Z}^+$ ,  $k \leq T$ , choose  $\theta^{(d)} \sim \text{Dirichlet}(\alpha_0 \mathbf{m}^{(k)})$ ; and

3. For each token  $i$  in document  $d$ ,

(a) choose a topic  $z_i \sim \text{Multinomial}(\theta^{(d)})$ , and

(b) choose a word  $w_i | z_i \sim \text{Multinomial}(\phi^{(z_i)})$ , where  $\phi^{(z)} \sim \text{Dirichlet}(\beta)$ .

To allow for any possible number of changepoints ranging from 0 to  $T - 1$ , we use Dirichlet process hidden Markov model (DP-HMM) that allows for one-step forward transitions through the discrete state space. Specifically, the transition matrix can be represented by

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & \cdots \\ 0 & p_{22} & p_{23} & \cdots \\ \vdots & \vdots & \ddots & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (6)$$

From state  $k$ , the only possible paths are to remain in state  $k$  or transition to the next state  $k + 1$  (i.e.,  $p_{kk} + p_{k,k+1} = 1$ ). We assume a transition matrix that only permits forward transitions to characterize the evolution of brand-related social media posts, as has been used to understand the evolution of household life cycles (e.g., Du and Kamakura 2006). With the Dirichlet process prior, the transition probability is given by

$$p(s_t = k | s_{t-1} = j) = \begin{cases} \frac{n_{jj} + v}{n_{jj} + v + \lambda} & k = j \\ \frac{\lambda}{n_{jj} + v + \lambda} & s_t \text{ takes a new state,} \end{cases} \quad (7)$$

where  $n_{jj}$  denotes the counts of transitions that have occurred so far from state  $k$  to itself,  $\lambda$  is the hyperprior that controls the tendency to remain in the current state, and  $v$  is the hyperprior that controls the tendency to transition to the next state (Ko et al. 2015; also see Appendix A.1 in the e-companion). Importantly, the DP-HMM does not require an a priori specification of the number of changepoints and samples only the states around the changepoints in each iteration (Ko et al. 2015). This makes it particularly well-suited for our examination of the topic evolution in social media conversations, as we may infer the number of conversational regimes from the data and need efficient sampling for a large amount of data. Although we adopt DP-HMM for this research, alternative specifications of multiple changepoints could work similarly (e.g., Barry and Hartigan 1993, Green 1995, Gopalakrishnan et al. 2017).

In addition to Dirichlet symmetric hyperprior  $\beta$  for the word distribution, we take the Dirichlet concentration parameter  $\alpha_0$  for topic sparsity as

predetermined (Griffiths and Steyvers 2004) but estimate the vector of probability measure  $\mathbf{m}^{(k)}$  (i.e., topic prevalence) for each regime  $k$ . Thus, the joint distribution of word  $w$  and topic assignment  $z$  can be represented by

$$P(w, z | \alpha_0 \mathbf{m}, \beta, s) = P(w | z, \beta) \prod_{t=1}^T \prod_{d \in \mathcal{C}_t} P(z^{(d)} | \alpha_0 \mathbf{m}^{(s_t)}), \quad (8)$$

where  $s$  denotes the sequence of state  $s_t$  at each time  $t$ ,  $z^{(d)}$  denotes the topic assignments of tokens in document  $d$ , and  $\mathcal{C}_t$  denotes the document collection at time  $t$ . We apply the following hybrid algorithm to draw topic assignments  $z$  by Gibbs sampling, estimate probability measure  $\mathbf{m}$  by fixed-point iteration, and draw state  $s$  by Gibbs sampling in turn. (See Appendix A.1 in the e-companion for details):

1.  $z | w, \mathbf{m}, s$ ,
2.  $\mathbf{m} | z, s$ , and
3.  $s | \mathbf{m}, z$ .

The proposed LDA-MLC model nests a number of cases that are worth noting. Our model nests the scenario in which there is a single regime (i.e.,  $s_1 = s_2 = \dots = s_T$ ). The resulting time-invariant model, consistent with that used by Puranam et al. (2017), differs from the standard Gibbs sampling procedure for LDA model (Griffiths and Steyvers 2004) in that we estimate the Dirichlet parameter governing the distribution of topics,  $\alpha_0 \mathbf{m}$ , from the data rather than assuming a diffuse prior. We refer to this as a modified LDA model.<sup>2</sup> In addition to nesting a static LDA model, our approach allows topics to flexibly evolve. If new topics were to emerge late in an observation period, this would manifest as a topic having a high prevalence only in the later regimes and a low (or 0) prevalence in earlier regimes. The LDA-MLC model also allows for topics to emerge and fade, with the conversation returning to its previous content. This would manifest as a regime shift (e.g., from regime 1 to regime 2) when the new topic emerges, followed by another regime shift (e.g., from regime 2 to regime 3) when the topic becomes less popular. In regime 3, we would observe topic proportions similar to that of regime 1, allowing us to conclude that the topic proportions have returned to their prior levels. We next describe the social media data we employ in our empirical analysis to demonstrate the predictive ability of the LDA-MLC model and how it may be used to detect conversational changes.

## Empirical Applications

### Volkswagen Emissions Testing Scandal

**Data.** To illustrate our LDA-MLC model, we first apply it to social media data pertaining to the Volkswagen automotive brand around the time of the emissions testing scandal, which arose in September 2015. We use

Crimson Hexagon, a popular social listening platform, to download the text of social media posts on blogs and discussion forums.<sup>3</sup> We construct a query to identify social media posts on blogs and discussion forums that contain the phrases “Volkswagen” or “VW” (case insensitive). Our goal is to demonstrate the LDA-MLC model’s ability to identify the shifts in the content of social media messages. We therefore pull comments before and after the news of the emissions testing scandal that broke on September 18, 2015, spanning a period from September 4, 2015 to October 1, 2015. We anticipate identifying a changepoint on September 18, 2015. As we will show, other changepoints are detected both before and after this date, demonstrating the model’s ability to capture both minor and major changepoints.

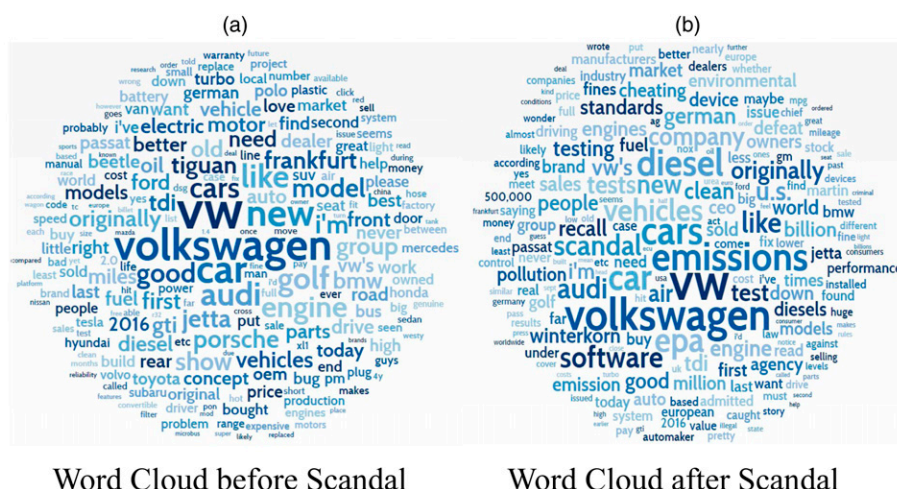
We preprocess the textual data by lowercasing and tokenizing characters and then removing punctuations (e.g., “/,” “@,” and “|”), numbers, stop words (e.g., “the,” “a,” and “and”), and URL links. In addition, we use the WordNet Lemmatizer (Miller 1995) to minimize the redundancy of words by taking their root form only (e.g., “cars” stems from “car,” and thus all occasions of “cars” and “car” are treated as the same). As a common practice in natural language processing, we use sparsity = 0.99 to cut off infrequently used words in the documents and provide the final set of words that will be analyzed. In doing so, we focus on the words that appear in at least 1% of the documents, consisting of a vocabulary with 1,081 words. This results in a total of 132,165 documents, from which we extract a 12.5% random sample.

Examining the volume of posts prior to the Volkswagen scandal on September 18, 2015, we see that the daily average during our observation period (2,086 posts/day) is considerably lower than that after news of the scandal broke (8,538 posts/day). Interestingly, the volume of posts reaches its peak not on the day of the initial news, but a few days later, between September 21 (11,335 posts) and September 24 (12,834 posts). Beyond fluctuations in volume, one would expect that the content of social media messages would differ before and after news of Volkswagen’s emission test cheating was released. Dividing our observation period into the days before and after the event, we present word clouds for the two periods in Figure 1. Figure 1(a) is generated from a random sample of 1,000 posts between September 11, 2015 and September 17, 2015, and Figure 1(b) is generated from a random sample of the same size between September 18, 2015 and September 24, 2015.

We see that posts before the scandal focused on Volkswagen models (e.g., “Jetta” and “Golf”), car brands (e.g., “Audi” and “Ford”), and product features (e.g., “engine” and “parts”). In sharp contrast, the posts after the scandal are dominated by the



**Figure 1.** (Color online) Comparison of Word Clouds for Volkswagen's Emissions Testing Scandal: (a) Word Cloud before Scandal, (b) Word Cloud after Scandal



scandal-related words (e.g., “emissions,” “diesel,” and “cheating”). Although Figure 1 provides model-free evidence for the changes in the content of social media posts, we can only produce these results knowing on what date to divide the data. We next present the results of the LDA-MLC model that can alert managers as to when such shifts occur.

**Model Comparison.** We estimate the model on the textual data of social media posts (i.e., forum, blogs, comments) that mention the brand between September 4, 2015 and October 1, 2015, during which the emissions testing scandal was reported. In contrast to studies that have used academic articles (Blei et al. 2003, Griffiths and Steyvers 2004) and product reviews (Tirunillai and Tellis 2014, Büschken and Allenby 2016) as the document corpus, social media conversations are characterized by highly diversified content surrounding brands (e.g., Schweidel and Moe 2014). We optimize the symmetric Dirichlet hyperprior for word distributions,  $\beta$ , with the data, as suggested by Wallach et al. (2009). The Dirichlet hyperprior  $\alpha_0 m$  works to smooth the count of tokens and a large value of  $\alpha_0$  would smooth out the topic proportions of short documents (Wallach et al. 2009). Given the short length of social media posts, we set the concentration parameter of Dirichlet hyperprior for topic distributions,  $\alpha_0 = 5$ . We increase the number of topics until the change of perplexity for the cross-validation sample flattens (Zhao et al. 2015), resulting in our use of 50 topics.<sup>4</sup> Perplexity is a widely used predictive metric in machine learning based on marginal likelihood (Blei et al. 2003, Grün and Hornik 2011).

To assess the performance of the LDA-MLC model, we use 80% of the social media posts of our sample within the one-month time window, which ranges

from September 4, 2015 to October 1, 2015 for calibration, the remaining 20% of our sample within the same time window for cross-validation, and the posts in the following five days from October 2, 2015 to October 6, 2015 for forecasting. We use the estimates of the modified LDA as the guidance of dispersed starting points to run multiple chains, average the log-marginal density (LMD) for model comparison, and record the estimate that yields the highest LMD (e.g., Blei et al. 2010). We compare our model with (1) the dynamic topic model (DTM) of Blei and Lafferty (2006) that assumes the hyperpriors of topic proportions and word distribution follow multivariate normal random walks, and (2) the modified LDA that assumes the hyperpriors of topic proportions are time-invariant (e.g., Puranam et al. 2017).

In Table 1, we summarize the model performance of the LDA-MLC model, the DTM, and the modified LDA model by reporting the log marginal likelihood of the calibration samples and the holdout samples. In conducting our holdout analysis, we focus on the model results obtained using 50 topics<sup>5</sup> and its prediction of the unseen data during the same time window as calibration as well as its forecasting for the social media posts from the following day (October 2, 2015), the following two days (October 2, 2015 to October 3, 2015), through the following five days (October 2, 2015 to October 6, 2015).

Although the DTM outperforms the LDA-MLC model and the modified LDA model with regard to the in-sample fit, the LDA-MLC model dominates both the DTM and the modified LDA in the cross-validation, suggesting that the DTM overfits the calibration data and performs poorer in predicting unseen data. In addition, whereas the DTM outperforms the LDA-MLC model in forecasting when considering social media posts from just the following day, which confirms the



**Table 1.** Model Comparison by LMD for Volkswagen Study

Study	Type	Date range	Modified LDA	LDA-MLC	DTM
Volkswagen	In-sample	9/4/2015–10/1/2015	–2,977,159.87	–2,978,832.72	–2,914,559.00
	Cross-validation	9/4/2015–10/1/2015	–330,833.25	–330,418.26	–343,997.00
	Forecasting	10/2/2015	–201,053.38	–201,068.93	–200,155.61
		10/2/2015–10/3/2015	–355,720.34	–355,706.26	–387,920.17
		10/2/2015–10/4/2015	–481,066.02	–481,128.18	–538,126.62
		10/2/2015–10/5/2015	–687,807.52	–687,944.56	–784,961.83
		10/2/2015–10/6/2015	–908,100.33	–908,307.77	–1,047,813.21

model comparison results of Blei and Lafferty (2006), both the LDA-MLC model and the modified LDA model outperform DTM in any longer date ranges. The LDA-MLC performs worse than the modified LDA on in-sample fit due to the latter model being more flexible (Büschken and Allenby 2016), and no better than the modified LDA on forecasting because of not knowing future regime shifts. We next present the detailed results of the LDA-MLC model applied to the Volkswagen social media data.

**Model Results.** The LDA-MLC model detects eight changepoints, or nine regimes, in the content of social media posts mentioning the Volkswagen brand during the one-month time window with a posterior mean probability in excess of 99.99%. We apply the relevance metric that balances the word distribution of topics with the word frequency in the text collection using weight  $\lambda = 0.6$  (Sievert and Shirley 2014). To illustrate the manner in which the topics shift, in Table 2 we present the five most prevalent topics in each regime (e.g., the topics corresponding to the five largest elements of  $m^{(k)}$  in regime  $k$ ), along with a selection of the most relevant words associated with the topics.<sup>6</sup>

As shown in Table 2, the two most prevalent topics are “Daily Activities,” which contains frequent words (e.g., “good,” “great,” and “love”) in daily conversations throughout the calibration time window. Prior empirical research applying the standard LDA model to social media posts has also identified topics that are comprised of words focused on daily activity and personal musings (e.g., Zhang et al. 2017). This topic accounts for approximately 19% of the content of the social media posts before September 18, 2015 in our data. In addition, the most prevalent topics before September 18, 2015 include vehicle usage (“User Experience,” which is marked by words such as “drive,” “buy,” and “own”) and maintenance (“Maintenance,” which is marked by words such as “oil,” “engine,” and “filter”).

The prevalence of “Daily Activities” drops to approximately 13% after the fifth changepoint that occurs at the beginning of September 18, 2015 on which the Environmental Protection Agency (EPA) accuses

Volkswagen of deliberately deploying “defeat devices” on 482,000 cars sold in the United State to mislead emission tests (Woodyard 2015). This event is captured by the topic “EPA Accusation” (marked by words such as “EPA,” “air,” and “defeat”) with a prevalence of 8.4%, and the topic “Fine” (characterized by words such as “fine,” “recall,” “issue,” and “pay”) with the prevalence of 7.0%. Thereafter, as shown in Table 2, the emissions scandal-related topics become more popular.

To characterize the nature of these changes in social media content, we identify those topics that experience the largest changes in prevalence (in magnitude) after each identified changepoint compared with its prevalence prior to the changepoint. We present the three topics that experience the largest increases and decreases, respectively, in prevalence and a sample of most relevant words of these topics in Table 3.

Before September 18, 2015, we identify three changepoints in Volkswagen-related social media posts. For instance, the prevalence of the topic “Concept Cars” (characterized by words including “Frankfurt,” “concept,” and “design”) increases by 1.8% after September 14, 2015, whereas the topic “Classics” (represented by words such as “bug,” “bus,” and “beetle”) decreases by 1.5%. Such changes in prevalence enable our model to identify the dates on which the content of social media conversations shift.

The fourth changepoint identified as September 18, 2015 is characterized by an increase in scandal-related topics “EPA Accusation” (+8.2%), “Fine” (+6.8%), and “EPA Accusation Details” (+4.1%). The content of the conversation changes a few days later after the fifth changepoint on September 21, 2015, with the “Scandal” (+4.7%), “Fuel” (+1.7%), and “Question” (+1.6%) increasing while the prevalence of “EPA Accusation” (–4.5%) subsides. This follows Volkswagen first admitting to cheating and formally apologizing (e.g., Woodyard 2015). The focus on the “Leadership” (+2.3%) topic begins to increase following the sixth changepoint on September 23, 2015, when a change of Volkswagen’s management occurred (Woodyard 2015). There is also increased discussion of the “German Economy” topic (+0.9%), consistent with the negative spillover across brands identified

**Table 2.** Most Relevant Topics in Each Regime of Volkswagen Study

Regime	Date range	Most prevalent topics	Topic label	Sample of most relevant words	Prevalence
1	9/4/2015–9/6/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	18.8%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	6.8%
		50	Maintenance	oil, engine, pump, filter, belt, timing, water, valve, fuel, pressure	5.7%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	5.5%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	5.5%
2	9/7/2015–9/8/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	20.1%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	6.8%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	6.7%
		20	Classics	bug, bus, van, beetle, original, paint, color, type, love, black	5.1%
		33	Family	city, family, trip, kid, town, summer, day, road, park, travel	4.8%
3	9/9/2015–9/13/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	18.7%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	7.2%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	5.8%
		50	Maintenance	oil, engine, pump, filter, belt, timing, water, valve, fuel, pressure	5.4%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	4.9%
4	9/14/2015–9/17/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	17.9%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	5.6%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	5.4%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	5.3%
		47	Package	rear, seat, wheel, steering, sport, speed, trim, design, interior, suspension	4.2%
5	9/18/2015–9/20/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	15.3%
		38	EPA Accusation	epa, air, clean, defeat, agency, software, device, recall, protection, california	8.4%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	7.0%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	5.1%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	4.8%
6	9/21/2015–9/22/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	12.8%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	7.7%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	6.5%
		46	Scandal	scandal, company, billion, german, carmaker, winterkorn, diesel, worldwide, software, euro	5.9%
		44	EPA Accusation Details	test, testing, emission, software, pass, cheat, mode, epa, real, road	5.0%

**Table 2.** (Continued)

Regime	Date range	Most prevalent topics	Topic label	Sample of most relevant words	Prevalence
7	9/23/2015–9/26/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	12.5%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	7.0%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	5.4%
		44	EPA Accusation Details	test, testing, emission, software, pass, cheat, mode, epa, real, device	4.8%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	4.7%
8	9/27/2015–9/29/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	12.7%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	6.7%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	5.8%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	4.7%
		46	Scandal	scandal, company, billion, german, carmaker, winterkorn, diesel, worldwide, software, euro	4.1%
9	9/30/2015–10/1/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	16.0%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	6.4%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	5.6%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	5.0%
		44	EPA Accusation Details	test, testing, emission, software, pass, cheat, mode, epa, real, device	3.3%

by Borah and Tellis (2016). With the seventh identified changepoint on September 27, 2015, “Warnings” (+1.7%) from Bosch about the abuse of their emissions testing software becomes popular, while the prevalence of other scandal-related topics like “Leadership” (−2.0%) and “Fine” (−1.9%) begin to diminish. On September 30, 2015, where the last changepoint is identified, “Daily Activities” (+3.3%) begins to increase in prevalence while the scandal-related topics such as “Warnings” (−1.5%) and “Scandal” (−2.0%) continue to decline.

Calculating the topic proportions for each document that is posted on a given day (as detailed in Equation (10) in Appendix A.1 in the e-companion), we illustrate how the average proportions of topics related to the emissions testing scandal shift over time in Figure 2(a) in contrast to the volume and sentiment proportions of posts provided by Crimson Hexagon in Figure 2(b).

The prevalence of topics related to the EPA’s accusation of Volkswagen’s cheating emerge on September 18, 2015. Although the observed negative sentiment increases with this change, the volume of posts does not soar until September 21, 2015. The

content of the social media posts continue to evolve, with the rapid emergence and gradual decline of breaking news (i.e., “EPA Accusation” and “Leadership” topics) giving way to an increase in the prevalence of details of the brand crisis (i.e., “EPA Accusation Details,” “Fuel,” and “Fine” topics) that linger longer. Although we observe these changes in the content of the social media posts, the negative sentiment increases initially and remains relatively constant after September 18, when the volume of posts gradually declines. As the comparison in Figure 2 reveals, the LDA-MLC models offer marketers deeper insight into how conversation surrounding their brands is changing over time compared with what the metrics reported in social listening platforms on their own will provide.

#### Changes in Topic Prevalence Versus the Contributor Base: Burger King and Under Armour

The Volkswagen empirical context demonstrates our model’s ability to identify changes in the topics discussed on social media. The dynamics observed in topic prevalence at the aggregate level may arise from two distinct sources. First, it may be that the topics

**Table 3.** Topics with Largest Increases and Decreases in Prevalence of Volkswagen Study

Changepoint	Date	Most shifted topics	Topic label	Sample of most relevant words	Prevalence change
1	9/7/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	1.3%
		27	Brake & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	1.3%
		45	Incidents	man, police, tell, white, left, road, day, wife, black, morning	1.1%
		31	Brands	audi, porsche, skoda, seat, group, luxury, bmw, brand, vag, bentley	−1.0%
		19	Durability	long, term, level, number, time, current, fact, high, potential, risk	−1.2%
		50	Maintenance	oil, engine, pump, filter, belt, timing, water, valve, fuel, pressure	−1.4%
2	9/9/2015	50	Maintenance	oil, engine, pump, filter, belt, timing, water, valve, fuel, pressure	1.1%
		25	Model	tdi, jetta, passat, golf, wagon, beetle, trade, sell, dealer, wife	0.8%
		31	Brands	audi, porsche, skoda, seat, group, luxury, bmw, brand, vag, bentley	0.7%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	−1.0%
		32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	−1.4%
		33	Family	city, family, trip, kid, town, summer, day, road, park, travel	−1.4%
3	9/14/2015	43	Concept Cars	frankfurt, concept, tiguan, suv, motor, bentley, design, version, renault, generation	1.8%
		48	Competition	sales, india, market, fiat, plant, automotive, brand, auto, year, chrysler	1.0%
		3	Electric Cars	electric, tesla, model, apple, car, battery, range, technology, project, charge	1.0%
		50	Maintenance	oil, engine, pump, filter, belt, timing, water, valve, fuel, pressure	−1.4%
		20	Classics	bug, bus, van, beetle, original, paint, color, type, love, black	−1.5%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	−1.8%
4	9/18/2015	38	EPA Accusation	epa, air, clean, defeat, agency, software, device, recall, protection, california	8.2%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	6.8%
		44	EPA Accusation Details	test, testing, emission, software, pass, cheat, mode, epa, real, device	4.1%
		47	Package	rear, seat, wheel, steering, sport, speed, trim, design, interior, suspension	−2.7%
		43	Concept Cars	frankfurt, concept, tiguan, suv, motor, bentley, design, version, renault, generation	−3.0%
		27	Brakes & Tires	part, kit, oem, brake, rear, installed, interior, arm, fit, bolt	−3.2%
5	9/21/2015	46	Scandal	scandal, company, billion, german, carmaker, winterkorn, diesel, worldwide, software, euro	4.7%
		15	Fuel	diesel, gasoline, clean, petrol, engine, fuel, powered, technology, dirty, emission	1.7%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	1.6%
		16	Forums	post, https, forum, thread, info, click, view, site, watch, link	−1.2%
		32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	−2.5%
		38	EPA Accusation	epa, air, clean, defeat, agency, software, device, recall, protection, california	−4.5%
6	9/23/2015	4	Leadership	ceo, board, winterkorn, executive, matthias, company, chief, chairman, porsche, resign	2.3%
		17	German Economy	german, germany, europe, industry, country, european, crisis, auto, euro, scandal	0.9%



**Table 3.** (Continued)

Changepoint	Date	Most shifted topics	Topic label	Sample of most relevant words	Prevalence change
7	9/27/2015	26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	0.8%
		38	EPA Accusation	epa, air, clean, defeat, agency, software, device, recall, protection, california	−1.2%
		46	Scandal	scandal, company, billion, german, carmaker, winterkorn, diesel, worldwide, software, euro	−1.5%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	−2.9%
		30	Warnings	bosch, software, illegal, vag, group, production, engineer, components, arm, reported	1.7%
		26	User Experience	car, drive, buy, own, reliable, year, owner, months, good, price	0.8%
		33	Family	city, family, trip, kid, town, summer, day, road, park, travel	0.7%
		38	EPA Accusation	epa, air, clean, defeat, agency, software, device, recall, protection, california	−1.2%
		42	Fine	fine, recall, epa, issue, wonder, pay, government, huge, owner, penalty	−1.9%
		4	Leadership	ceo, board, winterkorn, executive, matthias, company, chief, chairman, porsche, resign	−2.0%
8	9/30/2015	32	Daily Activities	good, lot, time, bit, pretty, better, work, great, love, stuff	3.3%
		16	Forums	post, https, forum, thread, info, click, view, site, watch, link	1.0%
		18	Service	service, dealer, warranty, dealership, customer, car, experience, excellent, extended, manager	0.9%
		22	Question	people, write, wrong, know, public, point, truth, happen, bad, trust	−1.1%
		30	Warnings	bosch, software, illegal, vag, group, production, engineer, components, arm, reported	−1.5%
		46	Scandal	scandal, company, billion, german, carmaker, winterkorn, diesel, worldwide, software, euro	−2.0%

mentioned by the same contributors shift over time. Second, the changes in topic prevalence observed in aggregate may arise not from changes in contributors' topic prevalences but from changes in the contributor base. That is, it is possible that a regime shift we detect at the aggregate level arises from new contributors who mention different topics from the contributors in the previous regime.

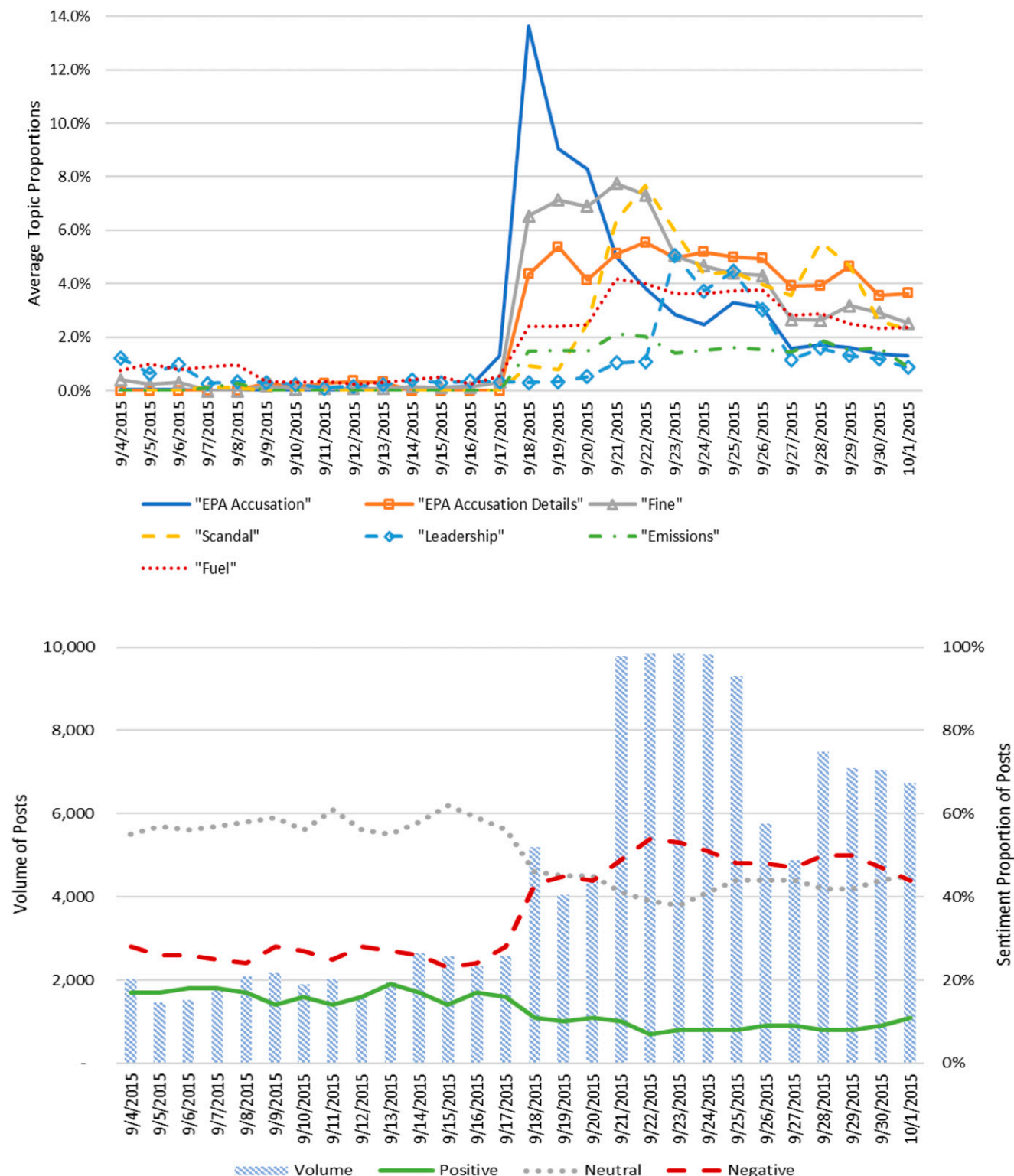
To investigate these alternative explanations for changes in topic prevalence, as well as demonstrate the proposed model's performance to brands that were not attracting the same level of attention as Volkswagen, we collected two data sets from Twitter: (1) 75,778 posts about Burger King from March 15, 2016 to April 11, 2016, during which time a new product (the Angriest Whopper) was released, and (2) 54,045 posts about Under Armour from March 15, 2018 to April 11, 2018, during which time a data breach was disclosed. We use Crimson Hexagon to pull the text of Twitter posts that contains case-insensitive keywords "Burger King" or "BurgerKing" for the former study and "Under Armour," "Under Armor," "UnderArmour," or "UnderArmor" for the

latter study. Using a sparsity of 0.99975 yields a vocabulary of 1,899 unique words for the Burger King study and 2,471 unique words for the Under Armour study. The usernames associated with Twitter posts enable us to identify the content contributed by each user over time, thus allowing us to distinguish between the two sources of observed dynamics. The Burger King investigation applies our modeling framework to a scenario in which the focal event (the release of a new product) is considerably less newsworthy than Volkswagen's emissions testing scandal and Under Armour's data breach.

The optimal numbers of topics for both studies are 60 with  $\alpha_0 = 1$  and optimized symmetric  $\beta$ . In both studies, we use 80% of all data for calibration and the remaining 20% for validation. As was the case in the Volkswagen analysis, the LDA-MLC model outperforms both the modified LDA and DTM in cross-validation in these empirical studies, as shown in Table 4.

We present the most shifted topics of these two studies at each changepoint in Appendix A.3 in the e-companion. Despite the smaller change in volume

**Figure 2.** (Color online) Changing Pattern of Scandal-Related Topics Versus Volume and Sentiment



Note. The number of posts that mention Volkswagen brand between day 18 and day 21 are censored due to the restrictions of Crimson Hexagon.

compared with the Volkswagen analysis, the LDA-MLC model accurately detects the focal events of Burger King's new product launch on March 29, 2016

and Under Armour's data breach on March 29, 2018. It also identifies other shifts such as the prank call that made Burger King workers break the store windows

**Table 4.** Model Comparison by LMD for Burger King and Under Armour Studies

Study	Type	Date range	Modified LDA	LDA-MLC	DTM
Burger King	In-sample	3/15/2016–4/11/2016	–1,383,374.54	–1,395,962.95	–1,218,600.46
	Cross-validation	3/15/2016–4/11/2016	–169,547.47	–167,733.95	–193,556.94
	Forecasting	4/12/2016	–53,354.20	–53,523.30	–49,037.72
		4/12/2016–4/13/2016	–130,591.39	–132,285.54	–161,933.53
		4/12/2016–4/14/2016	–190,524.48	–193,229.29	–248,243.47
		4/12/2016–4/15/2016	–242,864.67	–246,796.74	–324,928.10
		4/12/2016–4/16/2016	–275,617.18	–280,375.63	–375,376.14
Under Armour	In-sample	3/15/2018–4/11/2018	–1,587,204.26	–1,582,654.20	–1,503,786.58
	Cross-validation	3/15/2018–4/11/2018	–190,057.98	–188,115.45	–209,508.81
	Forecasting	4/12/2018	–77,303.23	–76,784.72	–75,493.82
		4/12/2018–4/13/2018	–133,305.33	–132,411.31	–164,648.93
		4/12/2018–4/14/2018	–188,782.11	–187,638.26	–252,769.92
		4/12/2018–4/15/2018	–260,952.30	–260,064.67	–348,013.12
		4/12/2018–4/16/2018	–315,654.70	–315,229.05	–416,875.46

on April 8, 2016 and the Under Armour All-America Camp's Orlando stop on March 25, 2018.

To examine the sources of the observed dynamics, we divide contributors into two groups by their posting behavior. One group consists of contributors who post both before and after the focal event (which we term “existing contributors”), and the other group consists of contributors who post either before the focal event (which we term “past contributors”) or after the focal event (which we term “new contributors”). Table 5 provides the distribution of contributors over the two groups.

The higher ratio of posts-to-authors from those contributors who post both before and after the focal event (3.67 posts per author for Burger King and 12.49 for Under Armour) compared with the ratio among those who only post before or after the focal event (1.14 posts per author for Burger King and 1.22 for Under Armour) suggests that the existing contributors are more engaged with the brand than the past and new contributors.

We plot the average topic proportions for the topics relevant to the focal events for each group of contributors and for all contributors in Figure 3, (a) and (b) in contrast to the volume and sentiment proportions of posts, respectively.

Comparing the content of posts from existing contributors to the new contributors who post after the launch of the Angriest Whopper, we see that they

exhibit similar topic proportions over time, except for “Red Buns.” This suggests that both changes in topic prevalence and changes in the contributor base contribute to the observed dynamics. In contrast to the Volkswagen analysis, we identify shifts in the content despite volume and sentiment remaining relatively stable throughout the observation period. In the case of Under Armour, however, we find that the existing contributors are much less likely to discuss the data breach than the new contributors who only post afterward. This suggests that changes in the contributor base are driving the dynamics observed to a large extent.

### Detecting Conversational Change-points

To illustrate the ability of the LDA-MLC model to detect conversational shifts in social media posts surrounding a brand, we estimate the model using a rolling window of one week throughout the observation period. We run the model with (1) no changepoint by setting the initial discrete state to one regime and (2) at most one changepoint by setting the initial state to two regimes, enabling us to ascertain if the addition of a changepoint is warranted by calculating the Bayes factor. In studying Volkswagen, for example, we estimate the LDA-MLC model on one week of data, shift the window one day forward and reestimate the model, repeating this process until the end of the rolling window occurs on October 7, 2015.

**Table 5.** User Distribution by Posting Behavior

Study	Group	Posting behavior	Number of authors	Number of posts
(a) Burger King	Past	Before only	25,131	28,370
	New	After only	31,345	36,190
	Existing	Both before and after	3,059	11,218
(b) Under Armour	Past	Before only	9,781	11,845
	New	After only	16,799	20,638
	Existing	Both before and after	1,726	21,562

We begin our analysis on August 22, 2015, yielding an additional week of data that we use to determine a reporting threshold that we apply to our observation period from September 4, 2015 to October 7, 2015 (e.g., Jeffreys 1961, p. 432). Based on our examination of the perplexity change for one week of data from the beginning of our observation period, the optimal

number of topics is 20 topics with  $\alpha_0 = 5$  in the study of Volkswagen and 40 topics with  $\alpha_0 = 1$  in the studies of Burger King and Under Armour, all using optimized symmetric  $\beta$ .

Each day of our observation horizon appears in seven rolling windows. We report in Figure 4 the number of rolling windows in which the start of a

**Figure 3.** (Color online) Average Topic Proportions of Event-Related Topics: (a) Burger King, (b) Under Armour

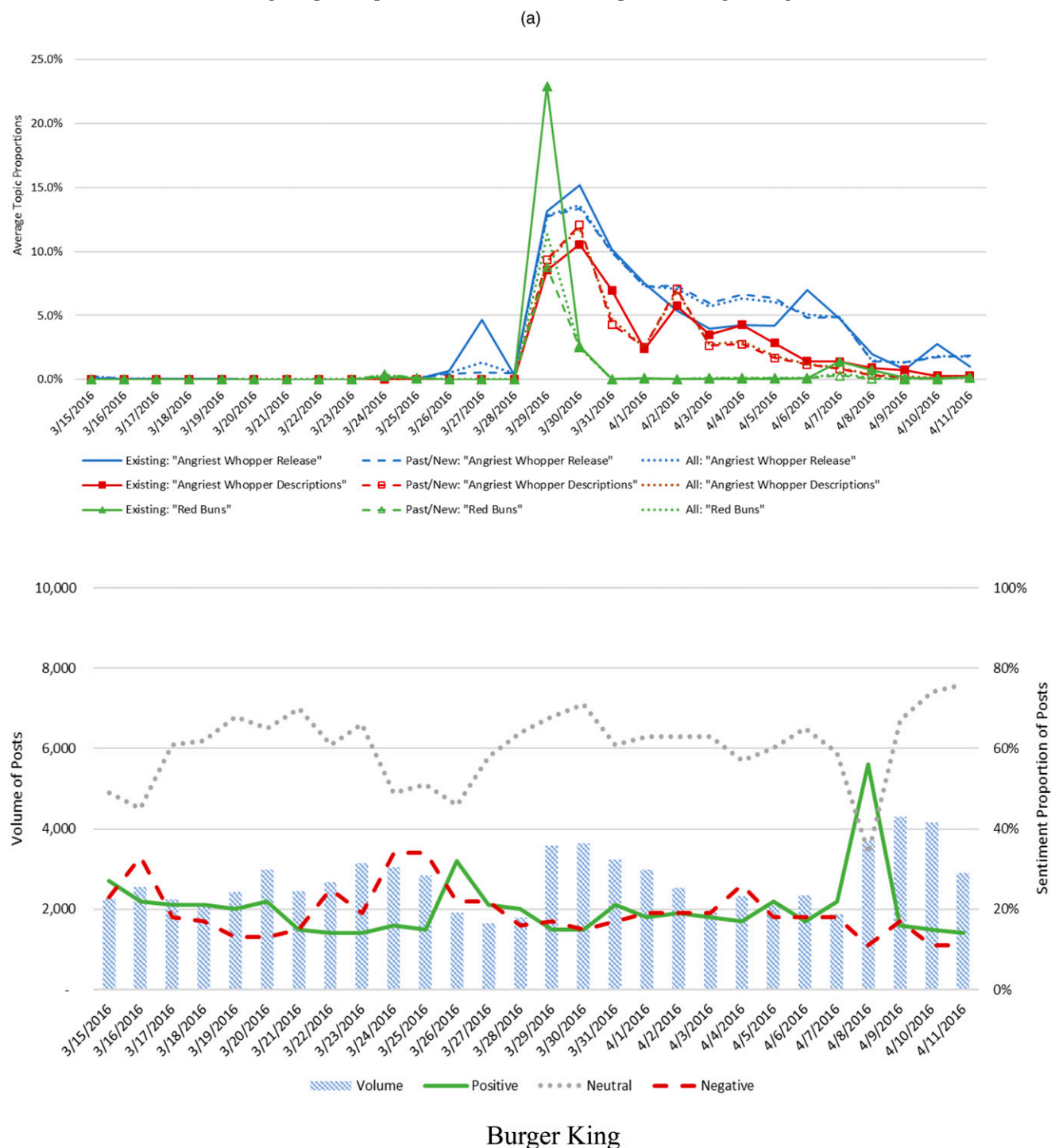
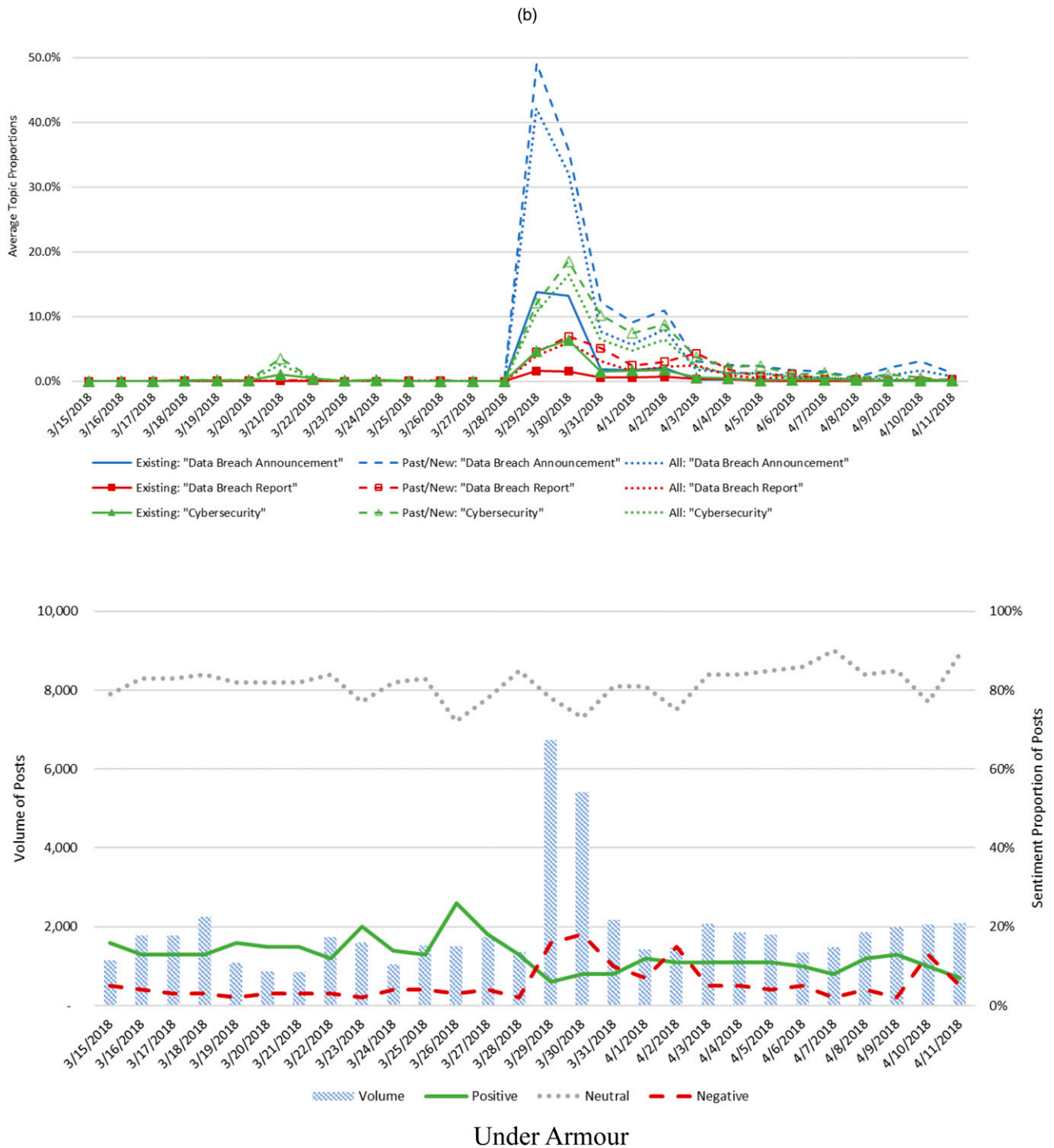




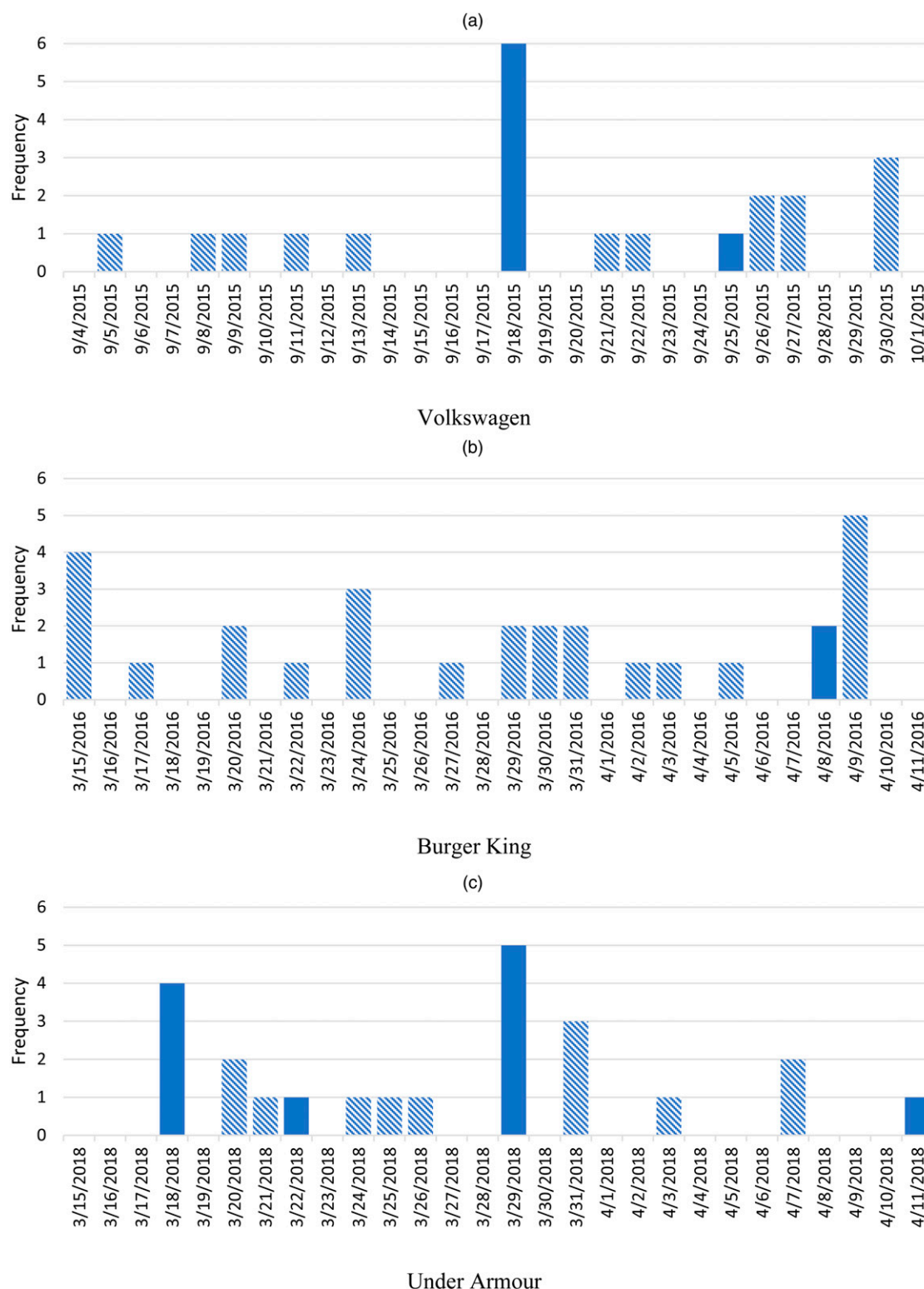
Figure 3. (Continued)



given date has the highest probability within the rolling window of being a changepoint. A changepoint is identified when the difference of the holdout sample LMD between the one changepoint model and no changepoint model exceeds a threshold, akin to setting the control limit in CuSum control charts (e.g.,

Granjon 2013). To identify this threshold for each brand, we calculate the interval of the difference of LMD between the one changepoint model and no changepoint model for the seven weekly rolling windows that precede our observation period. Based on these analyses, we set the threshold of difference

**Figure 4.** (Color online) Identified Changepoints in Rolling Window Analysis: (a) Volkswagen, (b) Burger King, (c) Under Armour



to 10 for Volkswagen, 50 for Burger King, and 50 for Under Armour.<sup>7</sup> As the changepoint is identified between days (e.g., between day 1 and 2, between day 2 and 3, etc.), this ranges from zero to six counts for a

given date. We also identify the dates that are recognized as the first postchangepoint day the first time they enter the rolling window. In the example of Volkswagen, for the week spanning September 12–18, 2015,

if the changepoint is identified as occurring at the end of September 17, then September 18 is the first post-changepoint day and is identified as the first time that it is included in a rolling window.

The solid bars indicate that the date is identified as the first postchangepoint day the first time it enters the rolling window. For such detection to occur, there must be a significant shift in the content of the social media posts on this day compared with the previous six days of the rolling window. For example, as shown in Figure 4, both the changepoint on September 18, 2015 in the Volkswagen study and the changepoint on March 29, 2018 in the Under Armour study are detected on the first day that it is included in a rolling window. The release date of the Angriest Whopper in the Burger King study is March 29, 2016. For the rolling window ending on March 29, 2016, the changepoint detected occurs on March 27, 2016, which is the first time that we observe an increase in average proportion of a topic related to the Angriest Whopper as shown in Figure 3(a) due to the posts about prerelease tasting. Although the LDA-MLC model can detect the changes in the social media conversations related to Volkswagen and Under Armour, which exhibit a spike in volume corresponding to the focal events, the analysis of Burger King suggests that a sufficient volume of conversation is necessary to detect the conversational shifts in a timely fashion.

## Discussion

In this research, we develop a topic model that discretely partitions social media posts based on the posting time to identify the shifts in the content of social media posts. We build upon the popular LDA modeling framework by incorporating a DP-HMM to allow for multiple latent changepoints, with topic prevalence varying before and after these changepoints. Our topic model with multiple latent changepoints allows us to capture a number of temporal patterns in conversation topics, including topic proportions remaining constant over time, temporary changes in topic prevalence that return to their prior levels, and topic proportions that continue to shift over time. Although there have been recent advances in the marketing literature in topic modeling (e.g., Tirunillai and Tellis 2014, Büschken and Allenby 2016, Liu and Toubia 2018), to the best of our knowledge we are among the first to incorporate the time at which social media content is contributed into a topic modeling framework.

As we illustrate, our modeling approach can be used by marketers to actively monitor social media conversations surrounding brands and identify when the content of these conversations shifts. By setting a reporting threshold based on historic data, our modeling framework can support marketers by alerting

them to significant changes in social media posts about the brand. This can provide them with an early indication of potential problems, such as consumer reactions to a new product or adverse reactions to new marketing campaigns. In addition to identifying conversational shifts as brands enter a crisis, it may also enable them to identify when they are emerging from a crisis based on changes in social media conversations.

Our research also highlights the importance of looking beyond common metrics in monitoring social media conversations. Although the volume and sentiment of posts is easily tracked and reported in social media listening tools, we show in our analysis of Burger King that changes in volume and sentiment need not occur for changes in the conversations to occur. Moreover, the shift in topics relating to Volkswagen's emission testing scandal occurs several days before the volume of social media posts peak.

By jointly monitoring shifts in structured data such as volume and sentiment, along with the unstructured content of social media posts, marketers can gain insight into what is driving a sudden influx of brand-related social media activity, whether it is a product launch (e.g., Dellarocas et al. 2007), advertising campaign (e.g., Fossen and Schweidel 2019) or customer service failure (e.g., Ma et al. 2015). Related to this, our research also informs managers as to which consumers are contributing posts that focus on different topics. We show that changes to the contributor base as well as shifts in the topics discussed by existing contributors contribute to the observed shifts in social media conversations. This decomposition of the aggregate comments into contributor groups can be crucial for marketers to understand. Contributions of brand-related social media posts may serve as a proxy for consumers' engagement with the brand (e.g., Malthouse et al. 2013). Differentiating between comments from consumers who may already be loyal to the brand and those who have only recently joined the conversation and may be less loyal to the brand will enable brands to assess how perceptions may differ between these groups and take appropriate actions. As an illustration, while Nike's advertising featuring Colin Kaepernick drew calls for a boycott on social media, online sales and the company's stock price surged.<sup>8</sup> Our analysis highlights the importance of identifying not only how the content of the conversation is changing, but also what is driving it, as this will inform marketers' next steps.

There are a number of ways in which the LDA-MLC modeling framework could be extended. From a methodological perspective, conventional LDA-based models ignore the ordering of words and semantic relations under the bag-of-words assumption. One way to relax such an assumption is to extend our model to an n-gram one. An alternative is to develop a model in

which the words are presented in vectors (e.g., Mikolov et al. 2013). It is also worthwhile to explore approaches that may reduce the computational burden. Although we rely on Markov Chain Monte Carlo (MCMC) to estimate the model, researchers using variational inference may be able to extend our work and apply it to brands that generate high volumes of conversation without the need for sampling (Blei et al. 2017). From a substantive perspective, if data were available for the same user across multiple social media platforms, our modeling approach could be extended to examine how users manage their online reputation based on the content contributed to particular platforms and the audience they intend to reach.

Whereas we incorporate dynamics in conversational content into the LDA-MLC model via latent changepoints, other factors may also contribute to conversation dynamics. For example, marketing efforts undertaken by a firm or its competitors may change the content that is discussed. Taking into account the interactions between marketing mix and online conversations would further tease apart the sources of the conversation dynamics. Given the interest in understanding the dynamics associated with structured user-generated content such as volume and sentiment, we hope that our work encourages future efforts to understand the dynamics of content and the potential applications of such methods for marketing insights.

## Endnotes

<sup>1</sup> See <http://www.slideshare.net/christinemoorman/the-cmo-survey-highlights-and-insights-feb-2016>.

<sup>2</sup> To ensure that the LDA-MLC model and the modified LDA model are identified, we conducted simulations using both models. The results of the simulation study are presented in Appendix A.2 in the e-companion. The results of our simulation show that the LDA-MLC model can detect abrupt changes in in topic prevalence, but that it is not suited for detecting more subtle shifts in topic prevalence.

<sup>3</sup> Due to the limits on the number of posts for which the full text can be exported from Twitter via Crimson Hexagon, we use data from blogs and discussion forums for which downloads are limited to 10,000 posts per day. When this limit is reached, which occurs for Volkswagen for the time period September 21–24, a random sample of 10,000 posts are collected and downloaded.

<sup>4</sup> Although we adopt unsupervised learning on Volkswagen-related text, one can always implement a semi-supervised model, such as by seeding specific words into pre-labeled topics if they know which topics are of interest a priori (e.g., Tirunillai and Tellis 2014, Puranam et al. 2017).

<sup>5</sup> Holdout performance was assessed varying the number of topics from 10 to 100 in increments of 10 topics. For all the number of topics considered, the relative performance of the three models is consistent, with the LDA-MLC model outperforming DTM, which in turn outperforms the modified LDA model.

<sup>6</sup> See Appendix A.4 in the e-companion for the prevalence of all the 50 topics over regimes and the top 20 most prevalent words associated with each topic.

<sup>7</sup> Higher reporting thresholds will result in fewer changepoints being identified and vice versa. Brands that are only interested in detecting major shifts may opt to set a higher reporting threshold, whereas brands may set lower reporting thresholds if they plan to manage social media conversations more actively.

<sup>8</sup> See <https://www.cnbc.com/2018/09/14/nikes-kaepernick-ad-should-fuel-sales-as-retailer-knows-its-consumer.html>.

## References

- Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.
- Barry D, Hartigan JA (1993) A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* 88(421):309–319.
- Blei DM, Lafferty JD (2006) Dynamic topic models. *Proc. 23rd Internat. Conf. Machine Learn.* (Association for Computing Machinery, Pittsburgh), 113–120.
- Blei DM, Griffiths TL, Jordan MI (2010) The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* 57(2):1–30.
- Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 518(112): 859–877.
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.
- Borah A, Tellis GJ (2016) Halo (spillover) effects in social media: Do product recalls of one brand hurt or help rival brands? *J. Marketing Res.* 53(2):143–160.
- Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Sci.* 35(6):831–998.
- Chib S (1998) Estimation and comparison of multiple change-point models. *J. Econometrics* 86(2):221–241.
- Culotta A, Cutler J (2016) Mining brand perceptions from Twitter social networks. *Marketing Sci.* 35(3):343–362.
- Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4):23–45.
- Du RY, Kamakura WA (2006) Household life cycles and lifestyles in the United States. *J. Marketing Res.* 43(1):121–132.
- Du RY, Kamakura WA (2012) Quantitative trendspotting. *J. Marketing Res.* 49(4):514–536.
- Fossen BL, Schweidel DA (2017) Television advertising and online word-of-mouth: An empirical investigation of social TV activity. *Marketing Sci.* 36(1):105–123.
- Fossen BL, Schweidel DA (2019) Social TV, advertising and sales: Are social shows good for advertisers? *Marketing Sci.* 38(2): 274–295.
- Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. *Marketing Sci.* 31(3):448–473.
- Gopalakrishnan A, Bradlow ET, Fader PS (2017) A cross-cohort changepoint model for customer-base analysis. *Marketing Sci.* 36(2):195–213.
- Granjon P (2013) The CuSum algorithm—A small review. Working paper, GIPSA-lab, Grenoble, France.
- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4): 711–732.
- Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(Suppl 1):5228–5235.
- Grün B, Hornik K (2011) Topicmodels: An R package for fitting topic models. *J. Statist. Software* 40(13):1–30.
- hipster (2015) Tdi owners possibly to be given free new cars. Deal of the century! Accessed December 19, 2019, <http://forums.redflagdeals.com/tdi-owners-possibly-given-free-new-cars-deal-century-1836797/#postcount23944519>.



- Jeffreys H (1961) *The Theory of Probability*, 3rd ed. (Oxford University Press, Oxford, UK).
- Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Sci.* 37(6):885–1052.
- Ko SIM, Chong TTL, Ghosh P (2015) Dirichlet process hidden Markov multiple change-point model. *Bayesian Anal.* 10(2):275–296.
- Kuksov D, Xie Y (2010) Pricing, frills, and customer ratings. *Marketing Sci.* 29(5):925–943.
- Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(October):881–894.
- Liu L, Dzyabura D, Mizik N (2017) Visual listening in: Extracting brand image portrayed on social media. Working paper, New York University, New York.
- Liu X, Singh PV, Srinivasan K (2016) A structured analysis of unstructured big data leveraging cloud computing. *Marketing Sci.* 35(3):363–388.
- Ma L, Sun B, Kekre S (2015) The squeaky wheel gets the grease—An empirical analysis of customer voice and firm intervention on Twitter. *Marketing Sci.* 34(5):627–645.
- Malthouse EC, Haenlein M, Skiera B, Wege E, Zhang M (2013) Managing customer relationships in the social media era: Introducing the social CRM house. *J. Interactive Marketing* 27(4):270–280.
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Preprint, submitted January 16, <https://arxiv.org/abs/1301.3781>.
- Miller GA (1995) WordNet: A lexical database for English. *Comm. ACM* 38(11):39–41.
- Moe WW, Schweidel DA (2012) Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* 31(3):372–386.
- Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.
- Nam H, Kannan PK (2014) Informational value of social tagging networks. *J. Marketing* 78(4):21–40.
- Nam H, Joshi YV, Kannan PK (2017) Harvesting brand information from social tags. *J. Marketing* 81(4):88–108.
- Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2):185–204.
- Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.
- Ngo S, Pilecki M (2016) The Forrester wave: Enterprise social listening platforms, Q1 2016. *Forrester* (March 2), Forrester Research, Cambridge, MA.
- Puranam D, Narayan V, Kadiyali V (2017) The effect of calorie posting regulation on consumer opinions: A flexible latent Dirichlet allocation model with informative priors. *Marketing Sci.* 36(5):726–746.
- Ringel DM, Skiera B (2016) Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Sci.* 35(3):511–534.
- Schweidel DA, Moe WW (2014) Listening in on social media: A joint model of sentiment and venue format choice. *J. Marketing Res.* 51(August):387–402.
- Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. *Management Sci.* 57(3):471–486.
- Sievert C, Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. *Proc. Workshop Interactive Language Learn. Visualization Interfaces* (Association for Computational Linguistics, Baltimore), 63–70.
- Sun M (2012) How does the variance of product ratings matter? *Management Sci.* 58(4):696–707.
- Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.
- Wallach HM, Mimno DM, McCallum A (2009) Rethinking LDA: Why priors matter. *Adv. Neural Inform. Processing Systems* 23: 1973–1981.
- Woodyard C (2015) Chronology: How VW's emissions scandal has mushroomed. *USA Today* (November 20), <http://www.usatoday.com/story/money/cars/2015/11/20/vw-volkswagen-chronology-emissions/76122812/>.
- Xiong G, Bharadwaj S (2014) Prerelease buzz evolution patterns and new product performance. *Marketing Sci.* 33(3):401–421.
- Zhang Y, Moe WW, Schweidel DA (2017) Modeling the role of message content and influencers in social media rebroadcasting. *Internat. J. Res. Marketing* 34(1):100–119.
- Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W (2015) A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* 16(Suppl 13):S8.