Toward a Comprehensive Understanding of User-Generated Content and Engagement Behavior on Facebook Business Pages

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Mochen Yang

IN PARTIAL FULFILLMENT OF THE REQUIERMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Yuqing Ren (Adviser)

Dr. Gediminas Adomavicius (Co-Adviser)

May 2018

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisers, Dr. Yuqing Ren and Dr. Gediminas Adomavicius. They have provided me with continuous support and guidance that helped me navigate through my Ph.D. study and made this dissertation possible. Their deep passion and immense expertise in research have been, and always will be, a constant source of motivation and inspiration for me. I could not have imagined having better advisers.

I would like to thank the rest of my committee members, Dr. Gordon Burtch and Dr. Jisu Huh, for their constructive feedback and comments. The insights and suggestions which they so patiently offered have tremendously improved the quality of this dissertation.

I thank all faculty members and doctoral students in the Department of Information and Decision Sciences, who have graciously guided and accompanied me during the past few years. I thank the Carlson School of Management for the generous financial support.

Last but not least, I am indebted to my wife, Wenqiuli Zhang. Without her unconditional support, encouragement, and love, I could not have completed my degree and this dissertation. I also thank my parents, Jiatian Yang and Yu Wang, for making me who I am.

Abstract

Social media platforms such as Facebook empower individual users to interact with companies and with each other on company-managed business pages. Users can generate content by posting directly to the business pages, and other users can engage with the content through multiple engagement features. Although such user-generated content (UGC in short) and associated engagement behaviors bear important consequences to the companies, they are not well understood. The three essays of my dissertation fill in this gap, by analyzing data collected from Facebook business pages with multiple empirical methods. The first essay examines the valence and content characteristics of user-generated posts on the Facebook business pages of multiple large companies across key consumer-oriented industries. It demonstrates that user posts on Facebook business pages represent a new form of UGC that is distinct from online product reviews generated by consumers, in terms of valence distribution and content types. Further, it highlights the important valence and content factors that influence two canonical types of engagement activities, i.e., liking and commenting. The second essay discusses how user engagement behaviors are shaped by engagement features on Facebook, and in particular, how the introduction of a new engagement feature affects the usage of existing features as well as overall engagement activities. It aims to uncover new insights regarding the interplay of multiple engagement features. Analyses show that, despite distinct functionalities, the usage of different features is not independent, and user posts that have received engagement are likely to obtain even more engagement of various types. The third essay addresses a methodological challenge of studying UGC on social media or other online

contexts, where researchers frequently seek to combine data mining with econometric modeling, but ignore the issue of measurement error and misclassification. Findings of my dissertation advance understanding of UGC and engagement behavior on social media brand pages, and have practical implications for social media platforms as well as businesses that have presence on these platforms.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Research Background

The increasingly pervasive use of social media technology has greatly transformed the way in which companies organize their online activities (Aral et al. 2013). In addition to delivering their messages through traditional, marketer-controlled communication channels, many businesses are increasingly relying on their digital presence on social media to build brand communities and engage their customers as fans (Goh et al. 2013; Dholakia and Durham 2010; Gallaugher and Ransbotham 2010).

One prominent example of businesses' digital presence on social media is Facebook business pages, which are dedicated brand pages managed by firms, for the purpose of engaging their customers as fans on Facebook. As of February 2017, there were more than 60 million business pages hosted on Facebook, and more than 1 billion visitors to such pages per month.[1] Facebook users who visit such business pages can participate in several types of activities, including (1) content *consumption*, i.e., viewing various content generated both by the businesses and by other users, (2) content *creation*, i.e., posting their own content in the form of user-generated posts, and (3) content *engagement*, i.e., engage with the content through many Facebook design features, including Like, Comment, or Share.

User engagement behaviors on Facebook business pages bear important consequences to the businesses. Recent studies in similar contexts have linked increased engagement with desired

---

[1] http://expandedramblings.com/index.php/facebook-page-statistics/. Last access 02/01/2017.

outcomes, such as growth in brand loyalty, purchase expenditures, and firm profitability (Dessart et al. 2015, Goh et al. 2013). Furthermore, engagement toward UGC (as opposed to marketer-generated content) deserves particular attention, because user-generated posts outweigh marketer-generated posts on Facebook business pages in both volume and impact on consumer purchase behaviors (Goh et al. 2013).

Despite the sheer number of business pages and the recognized importance of user engagement on these pages, our understanding of UGC and the associated user engagement behaviors on social media brand pages is quite limited. An extensive body of literature has been developed on online consumer reviews, a canonical form of UGC, in the contexts of online shopping websites and discussion forums (e.g., Chevalier and Mayzlin 2006; Forman et al. 2008; Godes and Mayzlin 2004, 2009; Archak et al. 2011). In fact, prior work studying UGC on social media brand pages has not conceptually differentiated such UGC from online consumer reviews (Goh et al. 2013; Ma et al. 2015). However, I find that user posts on social media brand pages represent a new form of UGC, which is conceptually different from online consumer reviews. Online reviews tend to be represented by feedback on specific products, in the format of ratings and textual descriptions, provided by customers (who typically have purchased the products) to inform other consumers' purchasing decisions. In comparison, user posts on social media brand pages are typically expressions of diverse topics, produced by any social media users who have an interest in interacting with the focal businesses or other users, and consumed by recipients with a wider variety of goals that are not necessarily purchase-oriented (Gallaugher and Ransbotham 2010;

Yang et al. 2014). Due to differences in content generation and consumption, previous knowledge on online consumer reviews may not be generalizable to UGC on social media brand pages. My dissertation research systematically investigates UGC and the associated user engagement behaviors in the context of Facebook business pages.

The dissertation follows the multi-essay format and consists of three essays addressing three inter-related research questions. The first essay explores *what kinds of content are produced on business pages, and which types of content attract user engagement*. The second essay examines *how engagement behaviors are shaped by specific design features of Facebook platform,* by leveraging a quasi-experiment opportunity on Facebook, i.e., the introduction of the "Reactions" feature. The third essay tackles a specific methodological challenge in studying UGC on social media as well as other online contexts, which discusses *how to leverage the combination of machine learning and econometric modeling to draw robust inferences*.

All three essays are related to the underlying theme of the dissertation. The first two essays make theoretical contributions and provide empirical evidence to the understanding of UGC and its associated engagement behavior on Facebook business pages, and the third essay makes methodological contributions to advance the robustness of inferences drawn from combining machine learning and econometric modeling, which is often used to study UGC in various online contexts. Each essay is written in a self-contained manner. It is important to note that the first two essays are currently under review, and the third essay has been published. The first two essays are collaborated with my adviser (Dr. Yuqing Ren) and co-adviser (Dr. Gediminas Adomavicius), and the third essay is collaborated with my adviser and co-adviser, as well as Dr. Gordon Burtch. To acknowledge their contributions, I use "we" throughout the dissertation when appropriate. The next section provides an overall of three individual essays.

1.2 Overview of Three Essays

My first essay examines the *valence and content characteristics* of user posts on Facebook business pages and the *antecedents* of engagement toward these posts. In particular, I study the *valence* (i.e., whether a post is positive or negative toward the focal business) and *content* (i.e., what a post is about) characteristics of the posts, and how post valence and content affect engagement, measured as the number of likes and comments received by a post.

While an extensive body of literature has studied online consumer reviews, i.e., a canonical example of online UGC, findings from this literature may not hold for user posts on Facebook business pages, because the two forms of UGC have several important distinctions. Online reviews tend to be structured feedback on specific products, in the format of ratings and textual descriptions, provided by customers, who typically have purchased the products, to inform other consumers' purchasing decisions. In comparison, UGC in brand communities on social media are typically open-ended expressions, provided by any users who have an interest in interacting with the businesses or other customers, and consumed by recipients with a wider variety of goals that are not necessarily purchase-oriented. These differences suggest that both what users post on social media platforms (e.g., Facebook business pages) and the resulting impact of the UGC are likely to be different from online reviews.

I analyzed a random sample of 12,000 user-generated posts from the business pages of 41 Fortune 500 companies in 6 consumer-oriented industries for the year 2012. These industries include Airlines, Commercial Banks, Consumer Products, Food and Drug Stores, General Merchandisers, and Specialty Retailers. I recruited Amazon Mechanical Turk workers to manually label the overall valence of each post as positive, negative, or neutral. To obtain the content categories of user posts, I took the grounded theory approach and followed an open coding stage and a structured coding stage. In the open coding stage, two research assistants blind to our research hypotheses independently analyzed a randomly selected set of 3,159 posts to identify common themes. Through several iterations, we came up with 7 categories: *positive testimonial and appreciation*, *complaint about product and service quality*, *complaint about money issues*, *complaint about corporate social responsibility (CSR) issues*, *customer question*, *customer suggestion*, and *irrelevant message*. Next, in the structured coding stage, I used Amazon Mechanical Turk to manually classify each post in the sample into the 7 categories. After labeling the valence and content of the 12,000 posts, I conducted a set of regression analyses, using negative binomial specification, to examine the relationships between post valence/content and the number of likes and comments a post received, while also controlling for several confounding factors, including post linguistic features, poster characteristics, post visibility, and company-level heterogeneity.

This essay has three key findings. First, across the 6 consumer-oriented industries, the valence of user-generated posts on Facebook business pages is overwhelmingly negative. All 6 industries had more negative posts than positive ones and, in particular, the Commercial Bank industry had 4 times more negative posts than positive posts. This pattern is in sharp contrast with the "J-shaped" distribution of online reviews on Amazon or other review sites, where positive reviews are the majority. Furthermore, negative posts on Facebook business pages also attract more likes and comments than positive posts on average. Second, at post content level, the three types of user complaints (i.e., complaints about quality of products/services, about monetary issues, or about CSR issues) exhibit heterogeneous relationships with different engagement outcomes. Quality- and money-related complaints tend to receive *fewer likes* but *more comments* than CSR-related complaints. Third, different engagement behaviors have different antecedents. For example, compared with neutral posts, positive posts tend to receive more likes but fewer comments, indicating that liking and commenting are used for different purposes.

My second essay focuses on the *dynamics* of engagement behavior, specifically, how the introduction of a new engagement feature changes the existing engagement patterns and the overall engagement activities with user-generated content. This question considers the interplay among multiple engagement features.

To answer this research question, I studied a feature change on Facebook as a quasi-experiment. On February 24th, 2016, Facebook introduced the "Reactions" feature across the entire platform. In addition to Liking and Commenting, users can also engage with individual posts by clicking one of the five "Reaction" buttons, including *love*, *haha*, *wow*, *sad*, and *angry*. This feature change provides a unique opportunity to study the effects of the Reaction buttons on Likes and

Comments. I collected all user posts on 29 Fortune 500 companies' business pages, generated 6 months before and 6 months after the introduction of the Reactions feature. I used a regression discontinuity specification to identify the overall effects of Reactions on Likes and Comments, by comparing the engagement activities of posts created immediately before the change with posts created immediately after the change. More importantly, to understand the impact of Reactions feature on user posts that have actually received Reactions, I relied on matching approach to create a matching sample of user posts prior to the feature change with comparable characteristics. I then examined if user posts with Reactions ended up receiving more or fewer Likes and Comments than they would have received before the feature change.

This essay has several key findings. Overall, the introduction of Reaction buttons increased the number of Likes but decreased the number of Comments received by an average user post on a Facebook business page. Meanwhile, the effect of Reactions exhibited interesting counter-intuitive heterogeneity. Specifically, I found that the introduction of Reactions feature had different effects on existing engagement features, depending on whether the post had received any Reactions. User posts that received at least one Reaction actually ended up receiving both more Likes and more Comments than they would have received before the feature change. In contrast, user posts that were created after the feature change yet did not receive any Reactions actually ended up receiving both fewer Likes and fewer Comments than they would have received before the feature change. In other words, the introduction of Reaction buttons heightened engagement intensity for content that received Reactions, yet cannibalized engagement for content that did not receive Reactions.

These effects, which were already detected within a month after the feature change, persisted after six months, suggesting long-term, lasting changes in engagement patterns.

My third essay focuses on a methodological challenge in studying UGC on social media or other online settings. Researchers frequently encounter the need to extract useful information from unstructured data such as texts and images, because UGC (e.g., Facebook posts) is typically presented in textual and graphical formats. To do so, researchers have sought to use predictive data mining techniques to generate new variables of interest. For example, data mining methods have been used to predict the valence of posts or tweets. The predicted variables (e.g., valence) are then added into explanatory models (e.g., regressions), usually as independent variables, to make statistical inferences.

Studies that have adopted such two-stage methodology are becoming increasingly prevalent in the IS discipline. A cursory search of recently published issues of top IS journals revealed at least 8 studies that have used this approach; I identified 6 recent studies in *Information Systems Research* (Gu et al. 2007, 2014; Aggarwal et al. 2012; Wang et al. 2013; Moreno and Terwiesch 2014; Singh et al. 2014) and 2 in *Management Science* (Archak et al. 2011; Lu et al. 2013). The most common application of data mining models in these studies was text classification that was used primarily for coding online user-generated content, such as consumer reviews and social media posts. The two-stage methodology has also been adopted in several fields outside the IS community, such as Marketing and Electronic Commerce (e.g., Tirunillai and Tellis 2012), Human-Computer Interaction and Decision Support Systems (e.g., Liu et al. 2012; Zhu et al. 2011, 2012), Economics (e.g., Jelveh et al. 2014), and Finance (see Fisher et al. 2016 for a review).

However, there is an important challenge with such two-stage methodology, rooted in the issue of measurement error. Because prediction is almost always imperfect, variables generated from the first stage data mining models inevitably contain errors. When the generated variables are added into the second stage econometric models, these errors manifest as measurement errors or misclassification in independent variables. If ignored, they can introduce systematic biases into estimations of the econometric models and threaten the validity of statistical inferences. This essay of my dissertation makes a methodological contribution by discussing the potential pitfalls of this popular research methodology and presenting approaches for mitigating them. It is related to the first two essays in that the methodological challenges and solutions discussed here can be applied to the empirical studies of UGC both on Facebook business pages and on other online platforms.

Specifically, I first use simulations to demonstrate that measurement error and misclassification, stemmed from predictive data mining models, can indeed introduce considerable biases into several commonly used econometric models, such as linear regressions, generalized linear regressions (e.g., Logit, Probit, and Poisson models). The biases can be severe even when the data mining models achieve relatively high predictive performance. For example, consider a binary classifier with 80% predictive precision on both classes; its predictions, when added into a simple linear regression as the sole binary independent variable, would result in a 40% underestimation of the coefficient for this variable.

Despite this challenging pitfall, I point out that, because standard data mining models provide measures of predictive performance that accurately quantify the errors, there is a unique opportunity of correcting the resulting biases. I review several existing methods for correcting the biases and focus on two specific methods that are particularly suitable in this context. The Simulation-Extrapolation (SIMEX) method applies to *continuous* variables with additive measurement error (Cook and Stefanski 1994), and the Misclassification-SIMEX method applies to *discrete* variables with misclassification (Küchenhoff et al. 2006). These methods have two key advantages over alternative error-correction methods: (1) they can be configured based solely on the performance indicators of data mining models (e.g., the confusion matrix or error variance measure), and (2) they can easily be applied to a variety of model specifications. Finally, I demonstrate the effectiveness of these two correction methods both with comprehensive simulation experiments and applications to three real world datasets, collected respectively from an online product review website, a social media platform, and an online crowdfunding website.

In reality, researchers often need to assess the magnitude and functional forms of errors in their data, before applying specific error correction approaches. To facilitate such assessment, I also develop a practical procedure for diagnosing errors and choosing error correction methods. Specifically, researchers can use the labeled dataset from the first-stage data mining model to diagnose the functional form of the error, the severity level of bias, and the effectiveness of correction methods, because both the true values and model-predicted values of the variables are observed. Equipped with knowledge from the diagnostic procedure, researchers can proceed to actual analyses using the unlabeled dataset and apply the chosen error-correction method.

Chapter 2. Valence and Content Characteristics of User Posts and Antecedents of User Engagement

2.1 Introduction

The increasingly pervasive use of social networking tools has greatly transformed the way in which companies organize their online marketing activities (Aral et al. 2013). In addition to delivering their messages through traditional, marketer-controlled communication channels, many businesses host brand communities on social networking platforms such as Facebook and Twitter, to engage their customers and encourage user-generated content (Goh et al. 2013; Dholakia and Durham 2010; Kiron et al. 2013). In particular, Facebook business page is a feature launched in 2007 to help businesses connect and interact with their customers. As of 2017, there have been more than 60 million business pages hosted on Facebook.[2] In some cases, customers become advocates who spread awareness and speak positively about the company's products and services (e.g., Swarovski's campaign on Facebook and Instagram encouraged customers to share photos of their products[3]). At the same time, challenges coexist with opportunities in managing UGC on social networking platforms. Companies usually have very little control over what customers post, and negative UGC can severely damage the brands (Goh et al. 2013).

Despite the enthusiasm and millions of dollars in investments from businesses, there have been limited theoretical understanding and empirical investigation of UGC in brand communities on social media (e.g., on Facebook business pages). Previous research on UGC has focused

---

[2] https://sproutsocial.com/insights/facebook-stats-for-marketers/. Last access 01/02/2018.

[3] https://www.facebook.com/business/success/swarovski. Last access 01/02/2018.

primarily on consumer reviews on online shopping websites and discussion forums around books, movies, TV shows, hotels, and restaurants (Chevalier and Mayzlin 2006; Forman et al. 2008; Godes and Mayzlin 2004, 2009; Archak et al. 2011; Ghose and Ipeirotis 2011). The few studies of UGC in brand communities on social media (e.g., Goh et al. 2013; Ma et al. 2015) have regarded the content as word-of-mouth and not conceptually differentiated such UGC from online consumer reviews. However, we believe that UGC in brand communities hosted on social media is conceptually different from online consumer reviews in several important ways. Online reviews tend to be structured feedback on specific products, in the format of ratings and textual descriptions, provided by customers, who typically have purchased the products, to inform other consumers' purchasing decisions. In comparison, UGC in brand communities on social media are typically open-ended expressions, provided by any users who have an interest in interacting with the businesses or other customers, and consumed by recipients with a wider variety of goals that are not necessarily purchase-oriented. These differences suggest that both what users post on social media platforms like Facebook business page and the resulting impact of the UGC are likely to be different from online reviews. To advance our understanding of this new form of UGC, we combine qualitative and quantitative analyses of archival data and insights from an exploratory online survey to answer two research questions. *(1) What kinds of posts, in terms of valence and content, do users generate on Facebook business pages? (2) How do posts' valence and content factors influence other users' engagement with the posts?*

We focus on user-generated posts ("user posts" in short) instead of marketer-generated posts because, compared to marketer-generated posts: (1) user posts are much larger in volume, and therefore can have a cumulatively greater impact; (2) they tend to be perceived as more credible, because peer customers are perceived to be more trustworthy than the company (Chen and Xie 2008); and (3) they have been shown to play a more influential role in driving purchases (Goh et al. 2013). Meanwhile, other researchers have studied the impact of marketer-generated posts on engagement and purchase behaviors through the lens of persuasive advertising (e.g., Goh et al. 2013; Lee et al. 2017). We believe that user posts on Facebook business pages is a distinctive phenomenon and warrants special attention.

In this essay, we focus on two post attributes: valence and content. Valence captures the degree to which a post is positive, negative, or neutral. Content captures the substance of a post, and can reflect the specific ways in which a post is positive, negative or neutral (e.g., whether it is a complaint about product quality or a complaint about corporate social responsibility issues while both being negative). Valence is a key characteristic that has been studied extensively in the online reviews literature (e.g., Godes and Mayzlin 2004). We decided to examine content, in addition to valence, because prior research has shown that the textual content of a message contains additional information that is often not captured by valence (e.g., Archak et al. 2011).

In terms of the impact of UGC, we study engagement behavior as the outcome for two reasons. First, increased engagement has been linked to increases in brand loyalty, purchase expenditures, and profitability (Dessart et al. 2015, Goh et al. 2013, Kim et al. 2013). Second, both theoretical and empirical understanding of engagement antecedents, especially in the context of social media, is still limited and represents a high-priority research direction (Maslowska et al. 2016, p. 470). In this essay, we examine two types of engagement behaviors: *liking* a post and *commenting* on a post, both of which are canonical ways in which users can engage with posts on Facebook, and both have been used to measure engagement in previous research of similar contexts (e.g. Lee et al. 2017; Gummerus et al. 2012). Different from prior research that often treats liking and commenting as interchangeable measures of engagement, we study liking and commenting as different forms of engagement behaviors with different cognitive costs, levels of interactivity, and antecedents.

Combining content analysis and econometric modeling, we analyzed 12,000 posts from the business pages of 41 Fortune 500 companies in 6 industries for the year 2012. In contrast to the widely observed positivity of online consumer reviews, users on Facebook business pages posted substantially more negative posts than positive ones. The ratio of negative to positive posts was 1.93 to 1. Econometric analyses showed that both positive and negative posts received more likes than neutral posts, and negative posts received more likes and more comments than positive posts. Analysis of post content revealed 7 salient categories as *positive testimonial and appreciation, complaint about product and service quality, complaint about money issue, complaint about social and environmental issues, customer question, customer suggestion,* and *irrelevant messages*. Our analysis also showed that the three types of complaints, while all being negative, received different numbers of likes and comments. Compared to *complaints about product and service quality* and *complaints about money issues, complaints about social and environmental issues* received more likes but fewer comments. Our results also confirmed that liking and commenting are two distinctive forms of engagement with different antecedents. Finally, we conducted an exploratory online survey to complement our analysis of the archival data. The survey provided valuable insights to help explain some of the key findings and advance our understanding of user motivations to visit the business page, contribute content, and engage with other users' posts.

Our work makes three novel contributions to the Information Systems literature. First, we are among the first to conceptually and empirically differentiate UGC in brand communities on social media from online consumer reviews, and to show how valence and content characteristics of UGC drive engagement in the new context. Our work establishes the predominantly negative valence as well as a content category framework for UGC in the context of Facebook business pages. Second, our research highlights the importance of examining specific content categories beyond valence. UGC with the same valence yet different content categories receive different types and levels of engagement. This finding has both theoretical implications on research of UGC and practical implications on social media marketing applications. Finally, our work advances the literature by highlighting the theoretical distinctions between liking and commenting as two different forms of engagement and empirically demonstrating how the same valence or content factors can have differential effects on the two.

2.2 Literature Review and Theoretical Development

Several bodies of literature in IS and Marketing shed light on our conceptualization and theorizing of user posts on Facebook business pages, including the literature on electronic word-of-mouth, online consumer reviews, and member engagement in online brand communities. In Sections 2.2.1 and 2.2.2, we draw insights from the electronic word-of-mouth and online review literature to theorize the likely valence and content characteristics of user posts (to discuss our first research question). In Sections 2.2.3 and 2.2.4, we summarize insights on engagement behaviors in online brand communities and theorize the impact of valence and content on engagement behavior (to discuss our second research question). Due to the relatively novel nature of our research context and lack of direct empirical evidence, we describe our speculations of the likely patterns, without explicitly formulating hypotheses.

2.2.1 Electronic Word-of-Mouth and Facebook Business Pages as the New Context

As a type of online UGC, user posts on Facebook business pages are closely related to electronic word-of-mouth. Word-of-mouth (WOM in short) refers to the informal communication by consumers to other consumers about their evaluations of goods and services (Anderson 1998) or about the ownership, usage, or characteristics of particular goods and services (Berger 2014). Existing literature on electronic WOM is primarily built upon studying online consumer reviews in various contexts. Online consumer reviews have emerged to become an influential force of consumer behavior, because the source (other customers) are perceived as more credible than the brand, and the channel (online, instead of offline) allows greater reach to the audience (Berger 2014). Several attributes of online consumer reviews, including volume, valence, and variance of review ratings have been associated with sales of a variety of products, such as sales of books (Chevalier and Mayzlin 2006), movie box office revenue (Liu 2006; Dellarocas et al. 2007; Duan et al. 2008), restaurant revenue (Lu et al. 2013), and sales of video games (Zhu and Zhang 2010).

In this essay, we argue that user posts on Facebook business pages are qualitatively different from online consumer reviews in several important ways such as *source*, *intended audience*, and potential *effects* on consumer behaviors, all of which are key dimensions of UGC (Berger 2014).

First of all, the two types of content are generated by different *sources*. While online consumer reviews such as product reviews on Amazon.com are typically generated by consumers with purchasing experiences, user posts on Facebook business pages can be generated by both consumers who had purchased products or services and Facebook users without purchasing experiences. In addition, the sources of online reviews and user posts may differ in their identifiability. While reviewer identity information is not always available for online reviews, user identity information is much more transparent and visible on social media platforms like Facebook. Identifiability is an important determinant of how recipients process the messages (Berger 2014). For example, online reviews containing identity-descriptive information were rated to be more helpful and associated with higher product sales than reviews without identity-descriptive information (Forman et al. 2008). Second, the two types of content have different *intended audiences*. For online consumer reviews, the intended audience is typically other consumers who are interested in purchasing the products. For user posts on Facebook business pages, the intended audience include both the companies and other Facebook users.[4] The difference in audience

[4] Characteristics of the source and intended audience of user posts on Facebook business pages are also confirmed by our exploratory online survey. We discuss the details of the survey in Section 2.5.

composition may influence what people choose to say (Berger 2014) and the degree to which the audience engages with the content. For example, compared to online reviews, posts on Facebook business pages may be more open-ended, in the sense that users can post not only information about a firms' products (Goh et al. 2013) but also complaints when customers perceive Facebook business pages as firms' "new" customer service centers (Kiron et al. 2013). Third, the *effects* of user posts on consumer behaviors are likely to be different from the effects of online reviews. While the readers of online consumer reviews often use the reviews to decide whether to buy a product, users on Facebook business pages may encounter a post at any stage of the marketing funnel (Anderson et al. 2011), such as awareness, consideration, or conversion. As a result, posts on Facebook business pages may not have as direct and pronounced an effect on purchase as online product reviews. Therefore, in this work, we focus on customer engagement as the outcome of interest, which, when properly cultivated, can act as a powerful driver of sales growth and profitability (Cvijikj and Michahelles 2013; Hoffman and Fodor 2010).

2.2.2 Valence and Content Characteristics of User Posts

The aforementioned distinctions between user posts and online consumer reviews have important implications on the valence and content characteristics of user posts. As a result, previous findings for online consumer reviews may not necessarily generalize to user posts on Facebook business pages.

A key observation of online consumer reviews is that their valence follows a "J-shaped" distribution, with large numbers of positive reviews, some negative reviews, and very few moderate ones (Hu et al. 2009). The fact that positive reviews typically outnumber negative reviews can be attributed to at least two reasons. First, people with high product valuations are more likely to purchase a product than people with low product valuations, and the former are also more likely to write positive reviews. Hu et al. (2009) refers to such behavior as "*purchasing bias*". Second, the abundance of positive reviews can be driven by consumers' *self-enhancement* motive, i.e., to look good to themselves and to others (Berger, 2014). Talking about positive experiences projects a more positive image of oneself (e.g., the person makes good choices or decisions) or serves as evidence of one's expertise (Wojnicki and Godes 2011). It can also boost the receiver's mood and make the audience feel better. In contrast, negative WOM may raise image impairment concerns, i.e., that the transmission of negative WOM may degrade one's image in the eyes of social others (Zhang et al. 2014).

However, for user posts on Facebook business pages, the purchasing bias and the self-enhancement tendency may not be as strong or prevalent. The purchasing bias is likely to become weaker because the source of user posts includes Facebook users with no purchasing experiences.[5] The self-enhancement tendency is also likely to be less prevalent because users are talking to a broader audience, not just other users. Posts on Facebook business pages are visible to both other users and the focal businesses, which makes the pages an effective channel to voice negative opinions, in order to punish a company that provide a bad product or service (Richins 1983; Sundaram et al. 1998), seek redress (Ma et al. 2015), or warn other consumers and help them avoid

[5] Our exploratory online survey (described in Section 2.5) also confirms that, among the users who had posted on Facebook business pages, about 10% reported they had never purchased products or services from the businesses.

bad experiences (van Doorn et al. 2010; Zhang et al. 2014). As a result, the valence distribution of user posts is not clear a priori, and we hope to characterize the pattern through our empirical analyses.

In terms of the content characteristic, online consumer reviews primarily focus on information and evaluations of products and services (Anderson 1998; Berger 2014). For other types of UGC in both offline and online settings, researchers have developed various taxonomies. Table 2.1 lists several frameworks that have been developed to represent content types in a variety of contexts.

Table 2.1. UGC Content Types in Different Contexts

Reference	Context	Content Types
Mangold et al. (1999)	Service marketplace	Quality, price, and value of service.
Richins and Root-Shaffer (1988)	Automobile purchase	Personal experience, advice-giving, product news, and negative WOM.
Schindler and Bickart (2012)	Online reviews for books and automobiles	Positive evaluative statements, negative evaluative statements, product-descriptive statements, and reviewer-descriptive statements.
Smith et al. (2012)	WOM on Twitter, Facebook, and YouTube	Promotional self-representation, brand-centric information, marketer-directed communication, response to online marketer action, factual brand information, and brand sentiment.
Cho et al. (2002)	Complaints in online feedback systems	Customer service, product quality, price, delivery problems, misleading information, trust issues, tracking, and promotion.

As shown in Table 2.1, the specific categories of UGC are highly context-specific. Although the literature does not provide a well-established framework to classify the content of user posts on Facebook business pages, it is likely that user posts may encompass a diverse set of content categories. For example, we have observed instances of user posts that are customer questions, suggestions, or complaints toward the businesses. Customer complaints has also been found in Ma et al. (2015) for brand-related tweets. In this research, we hope to develop a new content framework for user posts on Facebook business pages.

2.2.3 Customer Engagement in Online Brand Communities

In this section, we briefly review the literature on customer engagement in online brand communities to inform our theorizing about engagement on social media platforms. Engagement has been defined as "the intensity of an individual's participation and connection with the organization's offerings and activities initiative by either the customer or the organization" (Vivek et al. 2012, p. 4). Customer engagement plays a central role in the process of relational exchange in online brand communities where consumers interact with the brand and other consumers (McAlexander et al. 2002; Muniz and O'Guinn 2001). Brodie et al. (2013) define consumer engagement in an online brand community as "specific interactive experiences between consumers and the brand, and/or other members of the community" (p. 107). A key theme of Brodie et al.'s (2013) definition is that customer engagement is highly context-dependent, and its manifestations and levels of intensity can change over time and across contexts. Although customer engagement behaviors tend to have a primary focus on the brand, it can also focus on other targets such as UGC from other customers (Van Doorn et al. 2010). In particular, engagement behaviors in the context of an online brand community includes not only UGC creation (e.g., posting content) but also UGC consumption (e.g., liking and commenting on others' content) (Gummerus et al. 2012).

In this essay, we examine two types of engagement behaviors towards user posts: liking a post and commenting on a post. Liking and commenting are significant yet underexplored brand dialogue behaviors (Maslowska et al. 2016) through which customers can engage with the brand and other consumers. They add metavoicing or metaknowledge to user-generated content so that other users can gauge its popularity or value (Majchrack et al. 2013). Prior studies of engagement on Facebook business pages (e.g., Cvijikj and Michahelles 2013, Lee et al. 2017), while treating likes and comments as different dependent variables, typically did not try to *conceptually* differentiate these two engagement behaviors. In this essay, we conceptualize liking and commenting as two qualitatively different forms of engagement, and below we explain the theoretical reasoning behind the conceptualization.

Over the years, researchers have identified three key dimensions to characterize engagement behavior (Brodie et al. 2011), including the level of *cognitive* effort it requires or the amount of involvement it takes (Shevlin 2007; Oestreicher-Singer and Zalmanson 2013), the *emotional* states it expresses toward the target, and the *behavioral* manifestation it takes (Brodie et al. 2011). Drawing insights from the literature, we see at least two distinctions between liking and commenting: the level of effort or involvement and emotional complexity. Compared with commenting, liking is less cognitively demanding and represents a lower level of involvement with the content. More specifically, liking is a "lightweight, one-click feedback action" (Scissors et al. 2016), whereas commenting is a deliberate form of "composed communication" that takes time and cognitive capacity to compose (Burke and Kraut 2014; Swani et al. 2013). In terms of emotional complexity, liking is mainly used to express positive, affirmative emotions such as agreement, empathy, acceptance, or awareness (Scissors et al. 2016), whereas commenting can convey more complicated emotions such as appreciation, denial or disagreement, anger, or a combination of multiple emotions. We summarize the key differences between liking and commenting in Table 2.2.

Table 2.2. Conceptual Differences between Liking and Commenting as Engagement Behaviors towards User Posts

Reference	Applicable Dimensions	Liking a post	Commenting on a post
Van Doorn et al. (2010)	Valence	Mainly positive toward the post.	Could be positive or negative.
	Form/Modality	Requires low resource level.	Requires high resource level.
	Customer Goals	Express agreement, empathy, enjoyment, etc.	Express opinion and engage in discussion.
Brodie et al. (2011)	Cognitive	Requires low cognitive resources.	Requires high cognitive resources.
	Emotional	Mainly positive emotions.	Could be positive or negative.
	Behavioral	Liking and commenting are different engagement behaviors.	
Patterson et al. (2006)	Absorption	Low level of concentration on the post.	High level of concentration on the post.
	Dedication	Relatively weak involvement with the post.	Relatively strong sense of belonging to the post.
	Vigor	Low level of energy.	High level of energy.
	Interaction	Mainly one-way feedback.	Mainly two-way discussion.
Shevlin (2007)		Low level engagement.	High level engagement.
Oestreicher-Singer and Zalmanson (2013)		Mainly low-level behaviors such as *Content Consumption/Organization*.	Mainly high-level behavior such as *Community Involvement*.

2.2.4 Impact of Post Valence and Post Content on Engagement

Previous studies that investigated the antecedents of customer engagement on Facebook business pages have mostly focused on marketer-generated posts, i.e., how companies can engineer their posts to stimulate engagement behaviors and promote sales (e.g., Cvijikj and Michahelles 2013, Lee et al. 2017). In this section, we briefly review several related works on how valence and content characteristics of UGC affect engagement in the contexts of online reviews, online communities and social media platforms.

Two patterns emerged regarding the impact of message valence on readers' behaviors. First, compared to neutral messages, both positive and negative messages tend to have a greater impact. For example, studies of online reviews have shown that reviews with positive or negative valence have a greater impact on readers' perception of helpfulness and purchasing decisions than neutral ones (Yin et al. 2014; Chevalier and Mayzlin 2006). Corstjens and Umblijs (2012) found that positive posts on social media (e.g., Facebook) boosted sales of televisions, negative posts reduced sales, and neutral posts had no impact. Arguello et al. (2006) analyzed posts in several online communities and found that use of either positive or negative words in a message increased one's chance of getting a reply. Second, negative messages tend to have a stronger influence than positive messages, because negative information and emotions receive more processing and produce "larger, more consistent, more multifaceted, or more lasting effects" (Baumeister et al. 2001, p. 325). This is known as the "negativity bias", which has been confirmed in various contexts. For example, Chevalier and Mayzlin (2006) found 1-star (negative) reviews had a greater marginal impact on book sales than 5-star (positive) reviews. Arguello et al. (2006) showed that messages with more negative words were more likely to get a reply than messages with more positive words.

Compared to the impact of valence, studies on the impact of content are relatively sparse, although several studies have highlighted the importance of studying the textual content of UGC (e.g., Archak et al. 2011; Ghose and Ipeirotis 2011; Ghose et al. 2012). For example, Archak et al. (2011) showed that product features derived from textual reviews of cameras on Amazon, such as ease of use, product size, and picture quality, had a significant predictive power of product sales, over and above the volume and valence of the reviews. Also, there is evidence in the marketing literature that messages with the same valence but different types of content can have different effects on purchase intensions or other behaviors. Mohr and Webb (2005) found that, when subjects were presented with both corporate social responsibility (CSR, i.e., the firms' relationships with the environment or social welfare) and pricing information about a brand, negative CSR information decreased their purchase intent to a greater extent than negative pricing information.

Due to the differences between user posts on Facebook business pages and the other contexts, whether and how existing findings generalize to user posts remain unclear. Therefore, we rely on empirical analysis to uncover the relationships between post valence and content and engagement. We further speculate that the same valence/content factor may have different effects on likes and comments, because the two are distinct types of engagement. Compared to liking, commenting requires greater cognitive effort and is often used to convey more complicated emotions and opinions (Burke and Kraut 2014; Vries et al. 2012). Posts with negative valence thereby may attract more comments whereas posts with positive valence or other characteristics may attract more likes. Examining the antecedents of likes and comments separately will allow us to uncover these nuanced differences.

2.3 Methods

2.3.1 Research Setting and Data

We chose Facebook business pages as our research setting for several reasons. Facebook is the largest social media platform, both in terms of number of active users and the scale of marketing activities.[6] Its large user base and active interactions between businesses and users make it a suitable context to study our research questions. Figure 2.1 shows screenshots of Walmart's business page on Facebook and an example of user-generated posts. We built a software tool in Python to connect with Facebook Graph API to download data.

In this essay, we focus on Fortune 500 companies because they play an important role in the economy, are early adopters of Facebook business pages, and their business pages have reasonably high levels of traffic. Among all industries, we chose the 6 industries that are consumer-oriented, i.e., Airlines, Commercial Banks, Consumer Products, Food and Drug Stores, General Merchandisers, and Specialty Retailers, because the issue of UGC is much more relevant in consumer-oriented industries than others. We started with the Fortune 500 list from 2012 and found 41 companies in total that belonged to these 6 industries. We downloaded all posts on their Facebook pages in 2012, which were about 530 thousand in total. In addition to the textual content of each post, we also obtained data on creation time, media type (status, link, photo, or video), and number of likes and comments each post received. We then drew a stratified sample of 2,000 posts by company in each industry (i.e., the strata are companies within a specific industry), and obtained a sample of 12,000 posts. Table 2.3 lists the industries and corresponding companies.

2.3.2 Post Valence Analysis

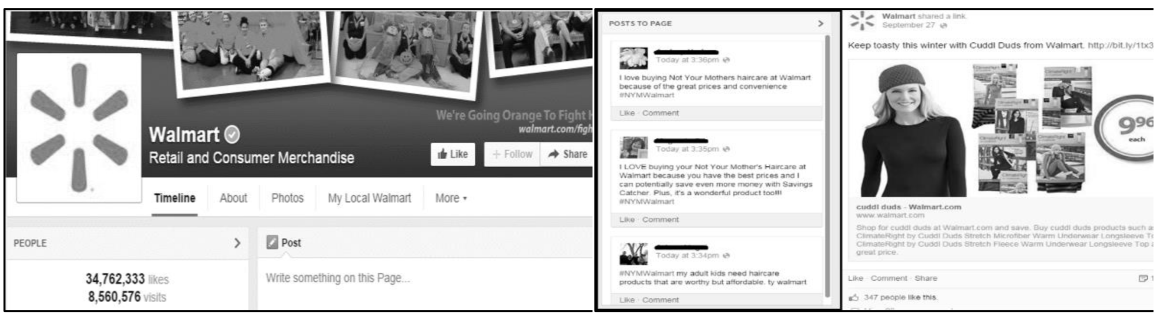
Our first analysis was to classify post valence. We recruited Amazon Mechanical Turk (MTurk)

---

[6] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/; https://www.huffingtonpost.com/young-entrepreneur-council/the-10-best-social-media_b_11654820.html. Last access 01/02/2018.

workers to perform this task. MTurk is an online marketplace for work where requesters can submit tasks, called Human Intelligent Tasks or HITs, to be completed at relatively low costs. Workers, also called Turkers, can accept a task, work on it, and get paid once their output is approved by requesters. Appendix 2.1 in the Online Supplement shows an example of our valence classification task.

Figure 2.1: Screenshots of the Facebook Business Page of Walmart and the Section of User-Generated Posts



(a) (b)

Note. Panel (a) shows Walmart's Facebook business page. Panel (b) shows the user posts, located in the "Posts to Page" section next to the main timeline.

Table 2.3. Data Collection: 6 Industries and Corresponding Companies

Industries	Companies
Airlines	Southwest Airlines, United, American Airlines, Delta, US Airways
Commercial Banks	Bank of America, Discover, Wells Fargo, U.S. Bank, Ally Bank, American Express, Sun Trust
Consumer Products	Dole, Kellogg's, Hershey's, Kraft Foods, Campbell's Soup, ConAgra Foods, PepsiCo, Land O'Lakes
Food and Drug Stores	Walgreens, CVS, Safeway, Rite Aid, Kroger
General Merchandisers	Target, Walmart, Macy's, Kohl's, Dollar General, Nordstrom, Dillard's, Sears, Family Dollar
Specialty Retailers	PetSmart, Best Buy, GameStop, Dick's Sporting Goods, AutoZone, Dollar Tree, Office Max

We instructed workers to carefully read the post and decide whether the post had an overall positive, negative, or neutral valence. To assure quality, we restricted the task to workers in the U.S.

who had a 95% or higher task acceptance rate. Each post was labeled by five workers, and we used the majority rule to determine the valence of a post. If three or more workers selected the same valence, then the post was labeled as having that valence. Using the majority rule, we were able to label 98.7% of the posts without ambiguity. For the remaining 1.3% where workers did not reach an agreement (e.g., 1 positive, 2 negative, 2 neutral votes), we tried two labeling strategies: (1) labeling them as neutral, or (2) labeling them as the relatively more dominant non-neutral valence (negative in the example). Our main results were qualitatively the same no matter which labeling strategy we used.[7] We presented results based on the first labeling strategy.

2.3.3 Post Content Analysis

Our second analysis was to classify post content. Due to the lack of established content framework for user posts on Facebook business pages, we took the Grounded Theory approach (Glaser and Strauss 1967), which is a qualitative approach to identify common themes and develop theory using empirical data. The approach includes two stages: *open coding* and *structured coding*.

During the *open coding* stage, two research assistants blind to the literature and our research questions independently analyzed a randomly selected set of 3,159 posts (not part of our 12,000 sample) to identify common themes. We worked with the two assistants through several iterations to make sure that the common themes had saturated and then started consolidating and organizing them into high-level categories. Our analysis suggested 7 categories: *positive*

---

[7] We ran additional analyses by (1) dropping the 1.3% of posts that lacked agreement in valence coding, (2) dropping the 3,899 posts that lacked unanimity in valence coding (i.e., not all 5 workers agreed on a single valence coding). Our main results remained qualitatively the same, confirming the robustness of our results.

testimonial and appreciation (*positive testimonial* in short), *complaint about product and service quality* (*quality complaint* in short), *complaint about money issues* (*money complaint* in short), *complaint about social and environmental issues* (*social complaint* in short), *customer question*, *customer suggestion*, and *irrelevant message*. Table 2.4 shows the definition and example of each content category.

During the *structured coding* stage, we first had the two research assistants code the sample of 3,159 posts into the 7 categories. We then posted this sample to MTurk with detailed instructions and illustrative examples to show how to classify the posts into the 7 categories. Appendix 2.2 shows an example of our content classification task. To assure quality, we restricted the work to workers in the U.S. who had "classification master" qualifications, meaning that they had consistently demonstrated high performance in classification tasks. Each post was labeled by five workers, and we used the majority rule to determine whether a post fell into a specific category. Across the 7 categories, Cohen's kappa ranged from 0.61 to 0.87 between the research assistants and MTurk coding. Having established the validity and reliability of using MTurk workers to do the classification, we posted the 12,000 posts, of which 80.3% were classified into one category, 9.75% into two categories, 0.53% into three or more categories, and 9.42% into no category. A post may fall into no category because its content was unusual, meaningless, or ambiguous that workers did not reach an agreement about any particular category.[8]

To assure the validity and generalizability of the content category framework, we

---

[8] E.g., "*I took southwest to Seattle*" [Southwest Airlines], "*Hi target*" [Target], and "*Special dark*" [Hershey's].

Table 2.4. Definitions and Examples of the 7 Content Categories

Categories	Definitions	Example Post
Positive Testimonial and Appreciation	The post includes a positive testimonial or a form of appreciation for the company (e.g., saying how wonderful the company is or how much the user loves it, thanking the company).	*Thanks for the amazing gift box! I cannot wait to try the cinnamon pops!!* [Kellogg's]
Complaint about Product and Service Quality	The post includes a complaint about product and service quality of the company (e.g., poor quality products or bad services).	*Not to be mean but Kelloge krave is one the worst tasting cereals I have eaten. I swear I wish I had my receipt or something.* [Kellogg's]
Complaint about Money Issues	The post includes a complaint about money issues with the company (e.g., hefty fees or high prices).	*Why do charge so much money for air fares in a city thats small in revenue? #corporatecrooks.*[Delta]
Complaint about Social and Environmental issues	The post includes a complaint about the company but it's NOT about product/service quality or money issues. Instead, it may be a complaint about the company's standing on social or environmental issues such as labor, human rights, social equality, or pollution.	*Chocolate is good, child labor is bad! Time to separate the two!!!!* [Hershey's]
Customer Question	The post includes a question directed at the company (e.g., inquiry about its products).	*My daughter just got diagnosed with a tree nut allergy- do you have a list of your products that are nut free? Thanks!* [Kellogg's]
Customer Suggestion	The post includes a customer suggestion to the company (e.g., recommendation of new products and service to offer).	*It would be really nice if the bags in the cereal boxes were resealable like zip-lock to keep the contents fresh.... just a suggestion.* [Kellogg's]
Irrelevant message	The post has nothing to do with the company on whose page this post appears. It may be user self-promotion, promotional links, adult content, etc.	*GOOD MORNING ERIKA CAN YOU BELIEVE SUMMER IS FADING AWAY FAST?* [Family Dollar]

Note. Corresponding company names on whose page the post was made are indicated in square brackets.

triangulated it with two other sources. First, we compared our framework with the content categories developed in previous literature, as summarized in Table 2.1. Some of our categories, such as *positive testimonial* and *quality/money complaint*, also appeared in previous frameworks (Mangold et al. 1999; Richins and Root-Shaffer 1988; Cho et al. 2002). Other categories, such as *social complaint* and *customer question/suggestion*, were unique to our context. This suggests that we have uncovered content categories that are both common in related contexts (e.g., offline WOM and online reviews) and unique to the context of Facebook business pages. Second, we surveyed practitioner articles on what users post on Facebook business pages and how business page owners should respond to user posts. The advice includes: providing customer support by answering customer questions, thanking and promoting positive testimonials from customers, and acknowledging customer suggestions or complaints.[9] These insights provide additional support and validation of our content category framework.

2.3.4 Variables

Our *dependent variable* is the engagement with a post, measured as the number of likes and the number of comments that a post received. Greater number of likes or comments indicates greater engagement. Our key independent variables are *post valence* and *post content categories*. For post valence, we created two dummy variables representing positive and negative valence. The three valence categories were mutually exclusive, and neutral valence served as the base. For content categories, we created seven dummy variables corresponding to the seven categories with posts that did not belong to any category as the base. We also included several control variables as

[9] Sources: http://www.socialmediaexaminer.com/social-media-research-shows-what-people-expect-from-brands; http://www.syncapse.com/why-consumers-become-facebook-brand-fans; http://www.verticalresponse.com/blog/5-facebook-no-nos-that-turn-off-your-customers. Last access 08/08/2016.

explained below.

Post Linguistic Characteristics. In addition to sentiment, several other linguistic features have been examined in prior literature, such as readability, vocabulary richness, and prototypicality (Johnson et al. 2015). Among them, message length and readability are two commonly studied features (Chevalier and Mayzlin 2006; Ghose and Ipeirotis 2011), and both can affect engagement. Longer messages tend to be more informative and include product specifics, which can reduce product quality uncertainty. As a result, readers often find longer messages more helpful or diagnostic than shorter messages (Mudambi and Schuff 2010). In online communities, researchers have found that longer messages are more likely to receive replies than shorter messages (Joyce and Kraut 2006). Similarly, readability has been shown to affect engagement in both online product reviews (Ghose and Ipeirotis 2011) and online communities (Arguello et al. 2006, Johnson et al. 2015). Messages that are easier to read and comprehend can be understood by more people and therefore attract greater engagement. In contrast, messages that use complex sentences and vocabularies are more difficult to understand and less likely to attract engagement or responses (Whittaker et al. 2003). We measured post length by the number of words in a post. We measured readability by Automated Readability Index[10] (ARI), which takes into account the average length of words and the average length of sentences (Smith and Senter 1967). Higher ARI score means the text has longer words or longer sentences and is written in a more sophisticated manner (Ghose and Ipeirotis 2011).

---

[10] ARI score = 4.71*(#characters/#words) + 0.5*(#words/#sentences) – 21.43

Poster Characteristics. Another factor that influences engagement with UGC is source characteristics, such as source credibility, network position, or participation patterns (Berger 2014). A key attribute of the source is activeness, which has been studied in prior literature. For example, Iyengar et al. (2011) studied the effect of social contagion on new medical product adoption among physicians. They found that heavy prescribers had a stronger impact on the adoption behavior of others than light prescribers. The disproportionate impact of active users may be attributed to their high involvement or their high status in the community (Godes 2011). In online communities, users who contribute more content and perform coordination activities tend to emerge as community leaders, and their activities become more visible and have a larger impact on other members (Preece and Shneiderman 2009). We proxied poster activeness by the total number of posts a given user posted in 2012 on the business page where the focal post appeared.

Post Context. The degree to which a message attracts attention and engagement also depends on contextual factors, such as when and where it is publicized. Reading and replying to messages take time and effort. The abundance of user-generated content on social media platforms implies that content published on the same platform will have to compete with other content for attention (Wang et al. 2013). In our study, we controlled for competition with three measures. First, at the page level, a busy and popular page has more traffic and therefore more intense competition for attention. We measured *page popularity* as the total number of posts posted on the business page in 2012, including both user-generated posts and marketer-generated posts. Second, at the individual post level, posts published within a short time window tend to compete with each other for attention. We measured *post-level user-generated content* (*post-level UGC* in short) and *post-level marketer-generated content* (*post-level MGC* in short) as the number of user- and marketer-generated posts that were posted from 24 hours before to 24 hours after a focal post was created on the page. Third, external factors unrelated to Facebook could also affect *general interest* in a company, and activities on the company's page on Facebook. For example, there was a spike of activities on Volkswagen's Facebook page after the revelation of its emission scandal.[11] To control for changes in the general interest in the company, we collected media reports from major newspapers and magazines using the LexisNexis database. We calculated the number of media reports about the company within 1 day prior to the creation of the focal post (denoted as *LexisNexis_1*).

Other Control Variables. We included several dummy variables to control for the industry of a company and the media type of a post (e.g., status, link, photo, and video). We also controlled for the size of the company by including company assets in 2012. Finally, we log-transformed 5 of our variables to reduce skewness including *word count*, *page popularity*, *post-level UGC/MGC*, and *asset*.

2.3.5 Data and Sample

From our initial sample of 12,000 posts, we excluded 4 sets of posts. First, we removed 174 posts that had fewer than 2 words or fewer than 6 characters because these posts did not contain enough meaningful information. Second, we removed 1,121 posts that were posted by third parties, such

---

[11] http://www.mediapost.com/publications/article/266265/reeling-vw-dials-social-activity-way-back.html. Last access 01/01/2018.

as non-profit organizations and local businesses, instead of individual users. As shown in Appendix 2.3, third-party posts were mostly *positive testimonial* and self-promotional messages (e.g., thanking the company for charity events) and differed substantially from individual user posts. Third, two companies, Land O'Lakes, Inc. and American Express, only had 3 user posts in total in our sample, which was insufficient for meaningful analysis. We removed these 3 posts. Finally, we identified and removed 21 posts with abnormal content, such as URLs or meaningless characters with no punctuation, because their valence and content could not be measured. After these removals, our sample included 10,681 user-generated posts from 39 companies.

2.3.6 Empirical Strategy

We ran negative binomial regression because our dependent variable was count data with significant overdispersion (supported by likelihood-ratio test for all regressions, $p < 0.001$). We considered both fixed effects and random effects models to control for company level heterogeneity. Because negative binomial regression is a nonlinear model, several issues need to be considered when choosing between fixed effects and random effects models. First, unconditional fixed effects model, with more than 20 dummy variables representing company-specific effects, is likely to produce inconsistent estimations (Hilbe 2011; Greene 2008) because of the incidental parameter problem (Neyman and Scott 1948). Alternatively, conditional fixed effects model (Hausman et al. 1984) provides consistent estimations, but has been criticized for failing to control for all company-invariant factors because the fixed effects are actually introduced into the overdispersion parameters (Allison and Waterman 2002). Although random effects model can provide consistent estimations, it relies on the assumption that company-specific effects are uncorrelated with other regressors (Greene 2008). Overall, there seems to be no clear consensus about the most appropriate approach. We chose conditional fixed effects and random effects negative binomial models because they generate consistent estimations, and ran them using the *xtnbreg* procedure in Stata.[12] The two models generated qualitatively similar results. Below we present results from the random effects negative binomial models, and include results from the conditional fixed effects models in Appendix 2.4.

We conducted several diagnostic analyses to check model assumptions. For multicollinearity, we ran regression models with OLS and checked variance inflation factors (VIF). All VIF values were below 4, suggesting multicollinearity was not a concern. We also checked for potential outliers. Residual plots showed 1 outlier post that received more than 1,000 likes. We removed it from our analysis (its inclusion did not change our results). We did not find any signs of heteroskedasticity issues. Table 2.5 lists our key variables and descriptive statistics. Correlation coefficients are included in Appendix 2.5.

One potential threat to the validity of our model is unobserved heterogeneity in post views. Users need to first encounter and view a post before liking it or commenting on it. Facebook is known to use an algorithm called "EdgeRank" to determine what posts appear in a user's *personal*

[12] Note that the random effects model estimated by *xtnbreg* procedure assumes the random effects follow a beta distribution (i.e., the conjugate prior of negative binomial distribution). Such specification allows a closed-form solution for the likelihood function and is therefore computationally preferred (Hilbe 2011). We also estimated a Gaussian-distributed random intercept model, but the results suggested the beta-distributed model had lower AIC and BIC, indicating he beta-distributed model had a better fit.

newsfeed and in what order. The algorithm considers affinity between the poster and the reader, the content of the post, and time decay since the creation of the content.[13] If the display of user posts on Facebook business page were subject to the influence of EdgeRank, then our regressions may suffer from omitted variable bias, because the unobserved heterogeneity in post views caused by Facebook's algorithm may be correlated with our independent variables and also affect our dependent variables. Such endogeneity would be very challenging to remove completely because the algorithm configuration is proprietary and unknown to the public.

However, the design of Facebook business pages differs from the design of personal newsfeed and allows us to alleviate the issue of endogeneity due to EdgeRank. User posts on a business page can be generated in one of two ways. Users can visit the page and write posts inside the "Post" textbox (as shown in Figure 2.1). Alternatively, users can publish posts on their own timelines and tag the business with the "@" sign (e.g., @WalMart). The posts that are directly created on a business page can only be seen by visitors to that page, and are not propagated to the posters' or other fans' friend networks. The visibility of these posts is not affected by the EdgeRank algorithm. In contrast, the posts with tags of the business will appear both on the business page and on the poster's own timelines. They can be seen by both visitors to the business page and the posters' friends, making these posts subject to the working of EdgeRank.[14] Another situation under which posts from a business page can be propagated to personal newsfeed is when the post gets *shared* by

[13] http://sproutsocial.com/insights/facebook-news-feed-algorithm-guide/. Last access 01/01/2018.

[14] Facebook does not reveal details about business page design. Authors acquired this information by opening a real business page on Facebook and experimenting with different ways of generating user posts.

Table 2.5. Variable Definitions and Descriptive Statistics (N = 10,640)

Category		Variables	Mean	SD	Min	Max
Dependent Variables		*Likes*	1.44	3.97	0	154
		Comments	1.67	3.56	0	81
Independent Variables	**Valence (dummy variables)**	*Positive Valence*	0.26	0.44	0	1
		Negative Valence	0.50	0.50	0	1
	Content Categories (dummy variables)	*Positive Testimonial*	0.23	0.42	0	1
		Quality Complaint	0.25	0.44	0	1
		Money Complaint	0.06	0.24	0	1
		Social Complaint	0.18	0.39	0	1
		Customer Question	0.19	0.39	0	1
		Customer Suggestion	0.07	0.26	0	1
		Irrelevant Message	0.08	0.27	0	1
Control Variables	**Post Linguistic Characteristics**	*Word Count*	44.46	66.90	2	1781
		ARI Score	5.03	4.94	−14.62	47.08
	Poster Characteristic	*User Activeness*	2.79	10.12	1	247
	Post Context	*Page Popularity* (in thousands)	19.78	24.55	1.16	125.86
		Post-Level UGC (in thousands)	0.39	1.26	0	10.27
		Post-Level MGC	7.98	25.18	0	534
		LexisNexis_1	5.58	8.71	0	81
	Industry (dummy variables)	*Airline*	0.18	0.38	0	1
		Commercial Bank	0.17	0.37	0	1
		Consumer Product	0.17	0.37	0	1
		Food and Drug Store	0.15	0.35	0	1
		General Merchandiser	0.17	0.38	0	1
		Specialty Retailer	0.17	0.38	0	1
	Media Type (dummy variables)	*Status*	0.94	0.24	0	1
		Link	0.02	0.15	0	1
		Photo	0.03	0.18	0	1
		Video	0.003[15]	0.06	0	1
	Assets (in billions)		270	648	2.33	2129

a user. Our sample included 32 posts with tags of the businesses and 8 posts that have received at

[15] We included posts with videos in our main analyses, but our major findings stayed unchanged after dropping 37 posts with videos (Appendix 2.12).

least 1 share.[16] The visibility of these 40 posts is partially determined by the EdgeRank algorithm, which we cannot completely control for. For the rest of the sample, we are confident that our *post context* variables (i.e., *page popularity*, *post-level UGC/MGC*, and *LexisNexis_1*) can sufficiently control for the heterogeneity of post views. After removing the 40 posts, our final sample included 10,640 posts. Notably, our results remained qualitatively the same even when these 40 posts were included (see Appendix 2.6).

2.4 Model Estimation and Results

In this section, we present the main empirical results. We discuss possible explanations and implications of our findings in later sections.

2.4.1 Distributions of Post Valence and Content

Comparison of our coding of post valence and post content categories showed both overlaps and discrepancies between the two. The majority of the posts that were labeled as having a positive valence were also classified into the positive content category (*positive testimonial*). The majority of the posts that were labeled as having a negative valence were also classified into the negative content categories (*quality complaint*, *money complaint*, or *social complaint*). Cohen's Kappa was 0.83 for positive and 0.89 for negative, indicating a high level of overlaps between valence and content. On the other hand, some content categories such as *customer question* and *customer suggestion* did not have a clear valence and could be positive, negative, or neutral. About 67% of *customer question* posts were labeled as neutral and 46% of *customer suggestion* posts were labeled

[16] These posts can be identified by examining the returned JSON objects from Facebook Graph API. Details can be found at https://developers.facebook.com/docs/graph-api/reference/v2.2/post.

as negative. Table 2.6 shows some examples of the two categories.

Of the 10,640 posts in our final sample, 5308 were negative and 2751 were positive, with the remaining 2581 being neutral. The ratio of negative to positive posts was 1.93 to 1. A chi-square test confirmed that negative posts were more prevalent than positive ones ($p < 0.001$). Furthermore, at the company level, 28 out of 39 companies had more negative posts than positive posts.[17] One-sided t-tests showed that, at both industry level and company level, there were significantly more negative posts than positive ones ($p < 0.01$ at industry level, $p < 0.001$ at company level). Figure 2.2a shows the percentage of positive and negative posts in the 6 industries. Negative posts were more prevalent than positive posts in every industry, with some variations across industries. For example, commercial banks had the highest percentage of negative posts, followed by airlines, consumer products, food and drug stores, specialty retailers, and general merchandisers. There were also differences across content categories. Figure 2.2b shows the percentages of the three types of complaints across industries. Airlines and commercial banks had higher levels of *quality complaint*, whereas consumer products companies had higher levels of *social complaint*. In general, *money complaint* was less common than *quality complaint* or *social complaint*.

2.4.2 Impact of Post Valence and Content on the Number of Likes

Table 2.7 shows the effects of post valence, content, and other variables on the number of likes. Model 1 included only control variables. Models 2, 3 and 4 incrementally added post linguistic features, post context, and poster characteristic. Model 5a added post valence and Model 5b added

[17] Companies that had more positive than negative posts are: Campbell's Condensed Soup, Discover, Dollar Tree, HERSHEY'S, Kraft Foods, Nordstrom, PepsiCo, PetSmart, Rite Aid, Sears Outlet Stores, and Southwest Airlines.

post content. Valence and content variables were not included in the same regression because they were highly correlated. We assessed model fit using Deviance, AIC, and BIC,[18] and the latter two adjusted for large samples and numbers of covariates (Raftery 1995). Models 5a and 5b had lowest BIC values and the best fit with our data. Therefore, we discuss the results of these two models.

Table 2.6. Customer Question and Customer Suggestion Posts with Different Valence

Content Categories	Valence	Example Posts
Customer Question	Positive	*I just saw that Campbells has a mobile truck in St.Louis! I wonder what they serve?? Soup only? I bet that truck would do great here in Vegas.* [Campbell's Soup]
	Negative	*I didn't receive my coupon :(what happened?* [Target]
	Neutral	*What are the movies this time? Anyone know yet?* [Best Buy]
Customer Suggestion	Positive	*I think u should add one more layer to the kit kat but make that peanut butter! I eat kit kats with P.B. OMG they are the best so how about adding one pb layer?* [Hershey's]
	Negative	*Remove Unsafe GMOs from your products!* [Kellogg's]
	Neutral	*Please create a Windows Phone app.* [Ally Bank]

Figure 2.2a: Percentages of Positive and Negative Posts across Industries

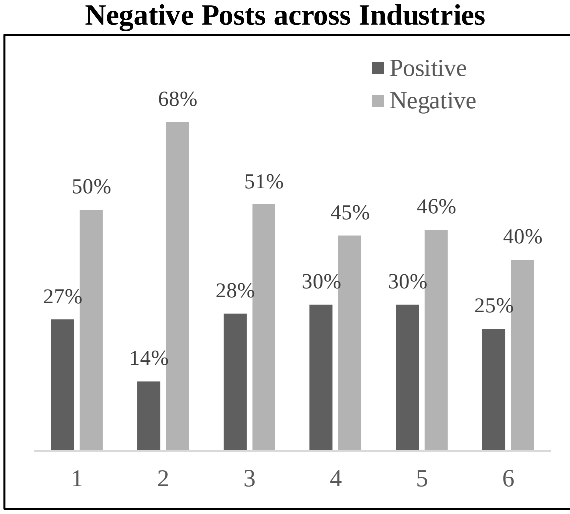
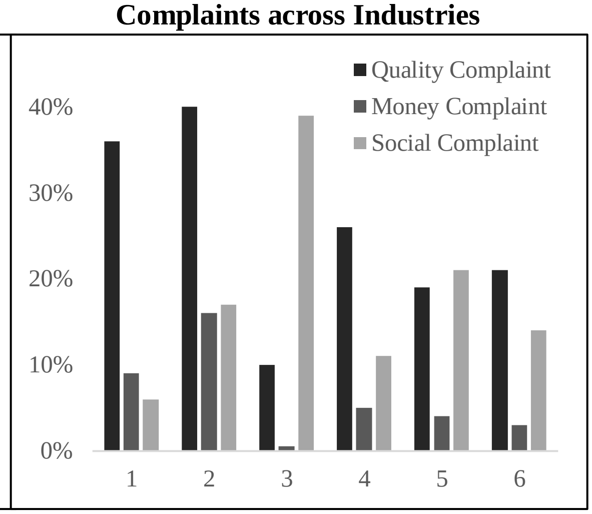


Figure 2.2b: Percentages of Different Types of Complaints across Industries



Note. Industry 1 – Airline; 2 – Commercial Banks; 3 – Consumer Products; 4 – Food and Drug Stores; 5 – General Merchandisers; 6 – Specialty Retailers.

[18] Denote the log-likelihood, degree of freedom, and sample size of estimated model as *LL*, *k*, and *N*, respectively. Then $Deviance = -2LL, AIC = 2k - 2LL, BIC = k\ln(N) - 2LL$.

As shown in Model 5a, compared to a post with neutral valence, a positive post received 72% more likes (b = 0.54, p < 0.001, *exp*(0.54) = 1.72) and a negative post received 118% more likes (b = 0.78, p < 0.001, *exp*(0.78) = 2.18). The coefficient of negative valence was significantly higher than the coefficient of positive valence (p < 0.001), suggesting that negative posts received more likes than positive posts. Furthermore, posts with the same valence but different content categories received different levels of likes. As shown in Model 5b, *social complaint* received more likes than *quality complaint* or *money complaint* (b = 0.93 versus 0.10 or 0.13, p < 0.001). All else being equal, *social complaint* received 129% more likes than *quality complaint* (*exp*(0.93-0.10) = 2.29), and 123% more likes than *money complaint* (*exp*(0.93-0.13) = 2.23). In addition, *customer suggestion* received more likes and *customer question* received fewer likes than posts not in any category. Posts that were irrelevant to the company's business did not differ from posts not belonging to any category.

Post linguistic features, post context, and poster characteristic also had significant effects on the number of likes. Both post length and readability were positively associated with the number of likes, suggesting that longer posts and more sophisticatedly written posts received more likes. We also explored the quadratic terms of word count and ARI score; neither was significant at the 0.05 level. Post context exhibited different effects on likes depending on which measure was used. At page level, posts on a popular page with higher traffic received fewer likes. At individual post level, being surrounded by more marketer-generated posts was associated with fewer likes, whereas being surrounded by more user-generated posts was associated with more likes. Higher general

Table 2.7. Random Effects Negative Binomial Regression on Likes (N = 10,640)

	Model 1	Model 2	Model 3	Model 4	Model 5a	Model 5b
Constant	0.13	−0.41	1.80***	1.75***	1.40***	1.43***
	(0.25)	(0.26)	(0.40)	(0.40)	(0.40)	(0.41)
Industry = Airlines	−0.23*	−0.27**	−0.34***	−0.35***	−0.39***	−0.25*
	(0.09)	(0.09)	(0.10)	(0.10)	(0.10)	(0.10)
Industry = Commercial Banks	−0.07	−0.12	−0.74**	−0.70**	−0.75**	−0.96***
	(0.18)	(0.18)	(0.24)	(0.24)	(0.25)	(0.25)
Industry = Consumer Products	0.06	0.05	−0.21	−0.20	−0.23*	−0.40***
	(0.09)	(0.09)	(0.11)	(0.11)	(0.11)	(0.12)
Industry = Food and Drug Stores	−0.67**	−0.69***	−0.72***	−0.72***	−0.74***	−0.80***
	(0.10)	(0.10)	(0.11)	(0.11)	(0.11)	(0.12)
Industry = General Merchandisers	−0.31**	−0.26**	−0.26**	−0.26**	−0.32**	−0.37***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Type = link	−0.55**	−0.54***	−0.48***	−0.49***	−0.34**	−0.46***
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Type = photo	0.75***	0.84***	0.88***	0.86***	0.97***	0.98***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)
Type = video	−0.12	0.01	0.13	0.10	0.24	0.15
	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)
Log(Asset)	−0.09**	−0.09**	−0.02	−0.02	−0.04	0.005
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Log(Word Count)		0.16***	0.17***	0.17***	0.10***	0.15***
		(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
ARI Score		0.01***	0.01***	0.01***	0.01***	0.01**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Log(Page Popularity)			−0.41***	−0.40***	−0.36***	−0.34***
			(0.05)	(0.05)	(0.05)	(0.05)
Log(Post-Level UGC)			0.26***	0.26***	0.22***	0.14***
			(0.01)	(0.01)	(0.01)	(0.01)
Log(Post-Level MGC)			−0.14***	−0.14***	−0.12***	−0.09***
			(0.02)	(0.02)	(0.02)	(0.02)
LexisNexis_1			0.01***	0.01***	0.01***	0.01***
			(0.00)	(0.00)	(0.00)	(0.00)
User Activeness				0.004***	0.004***	0.003**
				(0.00)	(0.00)	(0.00)
Positive Valence					0.54***	
					(0.05)	
Negative Valence					0.78***	
					(0.05)	
Positive Testimonial						0.25***
						(0.05)
Quality Complaint						0.10*
						(0.05)
Money Complaint						0.13
						(0.07)
Social Complaint						0.93***
						(0.05)
Customer Question						−0.54***
						(0.06)
Customer Suggestion						0.29***
						(0.05)
Irrelevant Message						−0.02
						(0.07)
Deviance	30250.22	30049.36	29540.22	29527.9	29225.58	28836.58
AIC	30274.22	30077.37	29576.22	29565.91	29267.58	28888.58
BIC	30361.49	30179.18	29707.12	29704.08	29420.30	29077.66

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses.

interest toward a focal company (*LexisNexis_1*) was associated with more likes. Finally, user activeness was positively associated with number of likes, suggesting that posts created by active users received more likes than those created by less active users. Our analyses also showed significant differences across industries and post media types. Posts on the pages of specialty retailers received more likes than posts on the pages of other industries. Compared to status updates, posts with links received fewer likes. Posts with photos received 163% more likes than text-only posts (b = 0.97, *exp*(0.97) = 2.63).

2.4.3 Impact of Post Valence and Content on the Number of Comments

Table 2.8 shows the effects of post valence, content, and other variables on the number of comments. Similarly, Models 5a and 5b had the lowest BIC values and the best fit with our data. We discuss the results of these two models. As shown in Model 5a, positive posts received 70% as many comments as neutral posts (b = -0.35, p < 0.001, *exp*(-0.35) = 0.70), and negative posts were not significantly different from neutral posts in the number of comments (b = 0.05, p = 0.14). However, compared with positive posts, negative posts received more comments (p < 0.001). Again, we found that posts with the same valence but different content categories received different levels of comments. As shown in Model 5b, *social complaint* received fewer comments than *quality complaint* or *money complaint* (b = -0.21 versus 0.27 or 0.13, p < 0.001). All else being equal, *social complaint* received 38% fewer comments than *quality complaint* (*exp*(-0.21-0.27) = 0.62), and 29% fewer comments than *money complaint* (*exp*(-0.21-0.13) = 0.71). In addition, our results suggested that, compared to posts not belonging to any categories, *customer question* received

Table 2.8. Random Effects Negative Binomial Regression on Comments (N = 10,640)

	Model 1	Model 2	Model 3	Model 4	Model 5a	Model 5b
Constant	−0.04	−0.77**	0.54	0.50	0.72	0.75
	(0.23)	(0.23)	(0.46)	(0.46)	(0.47)	(0.46)
Industry = Airlines	0.15	0.10	−0.12	−0.12	−0.06	−0.10
	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)	(0.09)
Industry = Commercial Banks	0.29*	0.32*	−0.88***	−0.86***	−0.77**	−0.50*
	(0.14)	(0.14)	(0.23)	(0.23)	(0.23)	(0.23)
Industry = Consumer Products	−0.49***	−0.45***	−0.92***	−0.91***	−0.86***	−0.74***
	(0.08)	(0.08)	(0.12)	(0.12)	(0.12)	(0.12)
Industry = Food and Drug Stores	0.25**	0.29***	−0.10	−0.10	−0.01	0.03
	(0.09)	(0.09)	(0.11)	(0.11)	(0.11)	(0.11)
Industry = General Merchandisers	−0.05	0.16	0.19*	0.19*	0.26**	0.25**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Type = link	−1.09***	−0.97***	−1.00***	−1.01***	−1.03***	−0.67***
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
Type = photo	−0.28***	−0.05	−0.01	−0.02	0.04	0.32***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Type = video	−1.24***	−1.04**	−1.01**	−1.03**	−1.07**	−0.66
	(0.35)	(0.35)	(0.35)	(0.35)	(0.35)	(0.35)
Log(Asset)	−0.05	−0.07**	0.07*	0.07	0.04	−0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Log(Word Count)		0.31***	0.29***	0.30***	0.26***	0.22***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ARI Score		−0.01**	−0.01*	−0.01*	−0.01**	−0.005
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Log(Page Popularity)			−0.16**	−0.15*	−0.13*	−0.10
			(0.06)	(0.06)	(0.06)	(0.06)
Log(Post-Level UGC)			−0.22***	−0.22***	−0.24***	−0.19***
			(0.01)	(0.01)	(0.01)	(0.02)
Log(Post-Level MGC)			0.08***	0.08***	0.08***	0.07***
			(0.02)	(0.02)	(0.02)	(0.02)
LexisNexis_1			0.0002	0.0001	−0.0002	0.002
			(0.00)	(0.00)	(0.00)	(0.00)
User Activeness				0.002*	0.002*	0.003**
				(0.00)	(0.00)	(0.00)
Positive Valence					−0.35***	
					(0.04)	
Negative Valence					0.05	
					(0.03)	
Positive Testimonial						−0.21***
						(0.04)
Quality Complaint						0.27***
						(0.04)
Money Complaint						0.13**
						(0.05)
Social Complaint						−0.21***
						(0.05)
Customer Question						0.36***
						(0.04)
Customer Suggestion						−0.05
						(0.05)
Irrelevant Message						−0.82***
						(0.08)
Deviance	34654.76	33996.5	33655.74	33651.78	33503.26	33121.1
AIC	34678.75	34024.50	33691.75	33689.78	33545.25	33173.11
BIC	34766.02	34126.31	33822.65	33827.95	33679.97	33362.19

Note. * $p<0.05$; ** $p<0.01$; *** $p<0.001$. Standard errors in parentheses.

more comments, *customer suggestion* was not significantly different, and *irrelevant message* received fewer comments (b = -0.82, p < 0.001).

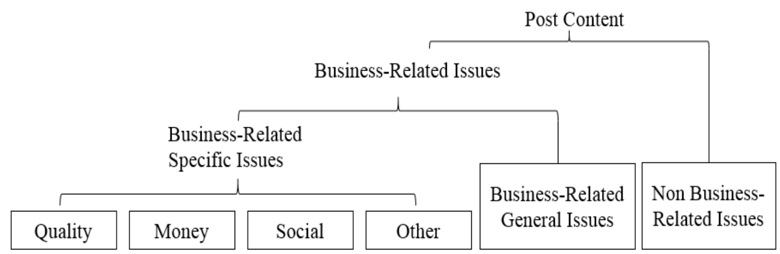
Post linguistic features, post context, and poster characteristic also had significant effects on the number of comments. According to Model 5b, longer posts received more comments, but ARI score was not significantly associated with the number of comments. The quadratic term of word count was significant (b = -0.04, p < 0.001) and the quadratic term of ARI score was not significant. Increasing post length from a few words to 150-400 words increased the number of comments, beyond which the effect began to decrease. Post context had different effects on comments depending on the measures. At individual post level, being surrounded by more marketer-generated posts was associated with more comments, and being surrounded by more user-generated posts was associated with fewer comments. Neither page popularity nor general interest toward a focal company was significant. Finally, user activeness was positively associated with the number of comments, suggesting that posts made by active users received more comments that those posted by less active users. Our analyses also showed significant differences across industries and post media types. Compared to specialty retailers, posts on the pages of commercial banks and consumer products companies received fewer comments and posts on the pages of general merchandisers received more comments. Posts with links received fewer comments than status updates, and posts with photos received more comments than status updates.

2.4.4 Analyses of Engagement with an Alternative Content Coding Scheme

The above results are based on a content coding scheme derived from the data-driven grounded theory approach. While the content categories represent what naturally emerged from an iterative open coding process (discussed in Section 2.3.3), they contained a mixture of several dimensions. For instance, the *quality complaint* category mixes the valence (negative) with the substance (statement about quality of products and services). Therefore, in this section, we repeated our main analyses under an alternative content coding scheme, where the content categories were regarded as being *orthogonal* to valence. Table 2.9 shows the alternative content coding scheme.

We considered 4 types of post valence as *positive*, *negative*, *neutral*, or *unclear*. Having an "unclear" valence means the valence of a post is ambiguous and cannot be determined (different from neutral valence). We considered 6 types of post content, based on whether the post is related to the focal business or not, and if so, whether the post is related to the *quality* of products and services, *money* issues, *social* issues, *other* specific business-related issues, or *general* business-related issues. The difference between the "*other*" content type and the "*general*" content type is that the former talks about specific aspects of the business that are not about quality, money, or social issues (e.g., "I want a job at Target"), whereas the latter talks about general aspects of the business without mentioning any specificity (e.g., "Macy's is a good place to shop"). Two research assistants helped code the valence and content of the posts as two *orthogonal* dimensions, giving rise to 4 (valence) × 6 (content), or 24, possible valence/content categories. Because a post may occasionally contain multiple different valence/content expressions (e.g., a post may talk positively about quality and negatively about money), we allowed each post to be coded in more than 1 of the 24 valence/content categories, and only less than 6% of posts had multiple labels. In addition, the research assistants also coded whether each post contained any *question* or *suggestion* toward the business, *independently* of post valence and content. Inter-rater reliability between the two research assistants was reasonably high on all major categories (Cohen's kappa between 0.6 and 0.8).

Table 2.9. Alternative Content Coding Scheme

<table>
<tr><th colspan="2">Orthogonal Dimensions</th><th>Coding Categories</th></tr>
<tr><td rowspan="2">Valence × Content Categories</td><td>Valence</td><td>Positive; Negative; Neutral; Unclear.</td></tr>
<tr><td>Content (actual categories are represented in boxes)</td><td>
</td></tr>
<tr><td colspan="2">Question</td><td>The post contains question toward the focal business: Yes or No</td></tr>
<tr><td colspan="2">Suggestion</td><td>The post contains suggestion toward the focal business: Yes or No</td></tr>
</table>

Analyses of the new coding revealed similar distributions of valence and content categories as what we reported in Section 2.4.1. About 52% of the posts are negative; 21% are positive and 19% are neutral. Second, posts coded as both negative and quality-related are most prevalent in Airlines (36%) and Commercial Banks (33%), and least prevalent in Consumer Products companies (9%). In comparison, posts coded as both negative and social-related are most prevalent in Consumer Products companies (42%) but least prevalent in Airlines (6%). Overall, descriptive patterns based on the new coding were consistent with earlier findings based on the content coding from the grounded-theory approach.

Next, we estimated a series of random effects negative binomial models to understand the impact of post valence and content, and the interaction of the two, on likes and comments respectively. The same set of control variables were included in these regressions. Tables 2.10 and

11 summarize the results. Note that we omitted coefficient estimates on all control variables for the sake of brevity.

Table 2.10. Random Effects Negative Binomial Regression on Likes (Alternative Content Coding Scheme)

	Model 1	Model 2	Model 3	Model 4	Model 5
Positive Valence	0.47***				
	(0.05)				
Negative Valence	0.64***				
	(0.05)				
Quality		−0.35***	−0.23***		
		(0.05)	(0.05)		
Money		−0.25***	−0.15***		
		(0.06)	(0.05)		
Social		0.70***	0.87***		
		(0.06)	(0.06)		
Other		−0.08			
		(0.07)			
Negative Quality				−0.32***	
				(0.05)	
Negative Money				−0.22***	
				(0.06)	
Negative Social				0.89***	
				(0.06)	
Negative Other				0.60***	
				(0.09)	
Question					−0.76***
					(0.06)
Suggestion					0.26***
					(0.04)
N	9,550	9,619	8,628	6,667	10,640
Sample Composition	Removed posts with unclear valence	Removed non-business-related posts	Removed non-business-related and general business-related posts	Removed non- business-related and general business-related posts. Removed neural and unclear posts	All posts
Base Comparison Group	Neural posts	General business-related posts	*Other* business-specific posts	*Positive* posts related to *quality*, *money*, *social*, and *other* issues	Posts that do not contain *question* and *suggestion*

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses.

Results in Tables 2.10 and 2.11 are qualitatively consistent with our earlier results, and also generate some new insights. According to Model 1 in both tables, compared to neutral posts, positive posts received more likes but fewer comments, and negative posts received both more likes and more comments. Furthermore, negative posts received both more likes and more comments than positive posts. Based on Model 2 and 3, under two different choices of comparison groups, posts about *social* issues received more likes but fewer comments than posts about *quality* or *money* issues. Moreover, as Model 4 indicated, negative posts about social issues (corresponding to *social complaints* in previous coding) received more likes but fewer comments than negative posts about quality and money issues (corresponding to *quality/money complaints* in previous coding). Finally, Model 5 showed that *questions* received fewer likes but more comments, whereas *suggestions* received more likes, than posts that did not contain questions and suggestions.

Table 2.11. Random Effects Negative Binomial Regression on Comments (Alternative Content Coding Scheme)

	Model 1	Model 2	Model 3	Model 4	Model 5
Positive Valence	−0.12***				
	(0.04)				
Negative Valence	0.24***				
	(0.04)				
Quality		0.21***	0.15***		
		(0.04)	(0.04)		
Money		0.20***	0.15***		
		(0.04)	(0.04)		
Social		−0.22***	−0.29***		
		(0.06)	(0.06)		
Other		−0.0002			
		(0.06)			
Negative Quality				0.35***	
				(0.04)	
Negative Money				0.28***	
				(0.04)	
Negative Social				−0.23***	
				(0.07)	
Negative Other				0.26***	
				(0.09)	
Question					0.45***
					(0.03)
Suggestion					−0.06
					(0.04)
N	9,550	9,619	8,628	6,667	10,640
Sample Composition	Removed posts with unclear valence	Removed non-business-related posts	Removed non-business-related and general business-related posts	Removed non- business-related and general business-related posts. Removed neural and unclear posts	All posts
Base Comparison Group	Neural posts	General business-related posts	*Other* business-specific posts	*Positive* posts related to *quality*, *money*, *social*, and *other* issues	Posts that do not contain *question* and *suggestion*

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses.

2.4.5 Robustness Checks

We ran several robustness checks with alternative operationalizations of our variables. First, we tried an alternative measure of *user activeness*, by calculating the number of posts a user made on a business page within the 3 months before the focal post. Because we only had data for 2012 and this new user activeness measure was not available for posts posted in January through March, we included posts that were posted between April and December. The vast majority of our findings were qualitatively the same, except for the significance of *quality complaint* (Appendix 2.7). Second, we considered several alternative measures of *post context*. We repeated the analyses by measuring page popularity as the total number of user-generated and marketer-generated posts within the 3 months prior to the focal post. All of our main results were qualitatively the same (Appendix 2.8). We also repeated our analyses by measuring *post-level UGC* and *post-level MGC* as the number of user-generated and marketer-generated posts only 24 hours before a focal post. The rationale is that including the number of posts posted after a focal post may cause simultaneity bias, because earlier posts may affect subsequent number of posts. With the new measure, our results were qualitatively the same (Appendix 2.9). Finally, we also repeated our analyses by changing the time window for the *general interest* variable from 1 day (*LexisNexis_1*) to 1 week or 2 weeks (*LexisNexis_7* or *LexisNexis_14*). All main results remained qualitatively the same (Appendix 2.10).

To further check the robustness of our findings, we estimated an alternative model specification where the dependent variable is binary, showing whether a post had received any likes or comments. This helped us address the concern that, if Facebook were to artificially boost the visibility of certain posts that had already received some likes or comments, these posts might gain even more engagement simply due to increased visibility. Such concern can be alleviated if we only study whether a post received any likes or comments, instead of the number of likes and comments. We estimated random effects logistic regressions, and our main findings were qualitatively the same (Appendix 2.11).

2.5 Exploratory Online Survey

Our objective in this essay is to examine what users post on Facebook business pages and the impact of post valence and content on engagement. Qualitative and quantitative analyses of archival data generated several insights and also raised some important questions. For example, who are the users who visit and post on Facebook business pages? What are their motivations of visiting the page, posting messages, and interacting with other users? What drives and explains the prevalence of negativity and the different antecedents of liking and commenting? We conducted an exploratory online survey to try to answer some of the questions and shed additional light on our key findings. We recruited participants from Amazon Mechanical Turk with two qualifications: (1) they must be in the U.S. and have Facebook accounts; and (2) they must have visited at least one business page on Facebook, and have read user posts on the page. We received a total of 123 valid responses. In the survey, we asked about (1) demographic information of users who have visited Facebook business pages including age, gender, and relationships with the businesses, and (2) motivations to visit business pages, read user posts, write posts, or like and comment on posts from other users using a five-point Likert scale. In designing the questions, we adapted established scales from relevant literature in online reviews and online communities (e.g., Hennig-Thurau and Walsh 2003; Hennig-Thurau et al. 2004; McAlexander et al. 2002), and created some new questions when we could not find established scales. A complete list of the survey questions is included in Appendix 2.13. Demographics of survey participants is shown in Table 2.12. A brief summary of the top motivations for visiting, reading, posting, liking and commenting on business pages of Fortune-500 companies is included in Appendix 2.14.

Table 2.12. Demographic Information of Survey Participants

Gender	Female: 58%; Male: 42%
Age	$\leq$ 25: 17%; 25-34: 47%; 35-44: 27%; 45-54: 5%; $\geq$ 55: 4%
Relationship with the focal businesses	- 70% have purchased products or services from the businesses; - 46% are considering purchasing from the businesses; - 10% have never purchased from the businesses before; - 4% are employees of the businesses.
Frequency of user activities	- $>$ 50% **visit** business pages at least once a week; - 70% **read** user posts at least monthly; - 73% have **posted** themselves at least once; - 84% have **liked** user posts at least once; - 76% have **commented** on user posts at least once.

Several things are worth noting from our survey responses. First, while the majority of visitors to Facebook business pages are customers with purchasing experiences, there are some visitors who have no purchasing experiences with the businesses. Second, users *visit* Facebook business pages and *read* user posts not only to get information about the companies' products and services and to learn about other users' experiences, but also for social reasons, e.g., being part of the user communities (agreed by 59% of participants). Third, the primary motivations for users to *post* on business pages include both sharing their experiences with other users, and requesting *customer service* from the businesses, by asking questions and making suggestions regarding the companies' products, services, or other issues (agreed by 55% of participants). Forth, the motivations for liking versus commenting are indeed different. While users *like* posts mainly because they agree with the posts or they share similar experiences with the posters, users *comment* on posts also to join the discussions by sharing their own experiences and to answer other users' questions.

To summarize, our survey responses confirmed our theoretical speculations that user-generated posts on Facebook business pages are conceptually different from online reviews. They are created by a combination of customers and users with no purchasing experiences. Their intended audience include both other users and the focal businesses. The motivations of creating and consuming user posts are not merely purchase-oriented and include a broad set such as requesting customer service and being part of the user community. As a relatively new platform for business-customer interactions, Facebook business pages seem to blend the elements of multiple phenomena, including but not limited to, electronic word-of-mouth among customers, online brand communities, and customer service interventions on social media.

Finally, our survey results regarding the motivations of creating and consuming user posts are also consistent with what prior literature has found. For example, Veirman et al. (2017) showed that users visit business pages ("lurking") because of the need for social interaction and entertainment, whereas they actively engage with the business pages, by posting their own messages and/or reacting to others' messages, mainly for social interaction and influencing other users and the focal brands (e.g., by giving suggestions to users and the brands). Similarly, Muntinga et al. (2017) found that the users create content on business pages in order to seek brand-related information, to engage in social interactions, as well as to influence other users and the brands.

2.6 Discussion

In this essay, we set out to answer two questions. What do users post on Facebook business pages? How do the valence and content of user posts affect engagement with the posts? We have at least three key findings that are worth discussing. First, we theorize and empirically demonstrate that user posts on Facebook business pages represent a relatively new phenomenon that is different from online consumer reviews. The prevalence of negative user posts is in sharp contrast with the "J-shaped" distribution of online reviews on Amazon or other sites, where positive reviews or 5-star ratings are the majority. We believe this contrast is partly driven by the differences in users' motivations to post on the two platforms. The primary motivation to write reviews on Amazon is to share one's opinions about the products and services and help other consumers' make better purchase decisions. In comparison, our exploratory online survey shows that users post on Facebook business pages to communicate with *both other users and the focal businesses*, around topics that are not necessarily related to the businesses' products or services. Most notably, the high volume of complaint messages and the additional customer questions and suggestions implies that some Facebook users regard business pages as a new channel to communicate directly with companies and expect companies to provide *customer service* on Facebook. From the businesses' perspective, this means that Facebook business pages are not just a channel for marketing activities, but also a channel to deliver customer services (Kiron et al. 2013) and manage customer relationships ("social CRM" discussed in Malthouse et al. 2013). The prevalence of negative messages on the business pages represents a significant challenge to many businesses. Future research should aim to uncover the complete nature of Facebook business pages as a new channel of interacting with customer and explore effective response strategies to manage customer complaints and other service requests on social media.

Our second finding is that liking and commenting are different types of engagement behaviors with distinct antecedents. Factors that increase the number of likes may not increase the number of comments. Empirical evidence from quantitative analysis and survey confirmed our conceptualization of liking and commenting as two distinctive forms of engagement behaviors. Compared to liking a post, commenting on a post requires more cognitive resources and involvement, and is typically used to express complicated emotions and opinions. Insights from our survey provides plausible explanations for the different antecedents to liking and commenting, particularly the finding that *social complaint* received more likes but fewer comments than *quality complaint* and *money complaint*. *Quality complaint* and *money complaint* generally pertain to personal experiences with the products and services, and therefore are likely to invite discussions and comments from other users, who may agree or disagree with the posters. This is consistent with the fact that many survey respondents rated “I want to add to the discussion by sharing my experience” as an important motivation of commenting on other users’ posts. *Social complaint* typically discusses social or environmental issues such as civil rights, child labor, and pollution that have broad social appeal. Liking such posts expresses agreement and empathy with the posters. This is consistent with our observation that survey respondents rated “I agree with the content of the posts" as the most important reason for liking.

Our third finding is the interplay between post valence and post content, and how going beyond valence to study the impact of post content reveals interesting heterogeneity among different kinds of posts. Notably, while the three types of customer complaints (respectively about *quality*, *money*, and *social* issues) are all negative in valence, we found they have different effects on engagement. As discussed in Section 2.4.4, compared to positive posts, *quality complaint* and *money complaint* received fewer likes and more comments whereas *social complaint* received more likes and fewer comments. We would have missed these nuanced but important effects had we only examined valence without differentiating the various ways in which a message could be negative. These findings demonstrate the benefit of combining sentiment analysis and content analysis to obtain deeper insights from textual data. Besides the grounded theory approach we adopted in our study, in future work, researchers should also consider alternative data-driven methods such as topic modeling (Blei 2012) and use it to complement human coding, to discover common themes and content categories in textual data.

Our research has important implications for social media marketing practice. First, companies have little control over how users behave and what users post on their business pages. This is of particular concern to companies whose users are more likely to use the page as an outlet to complain and vent their negative feelings. Companies should be aware of this challenge and not simply regard Facebook business page as a marketing channel. Instead, companies should carefully consider and evaluate whether Facebook business pages is an effective venue to interact with their customers and how prepared they are in managing possible user posts, especially the negative ones. Negative voices should not be left unattended, because they tend to attract more engagement than positive and neutral ones. Instead, they may reflect potential or pervasive issues of the companies' products, services, and corporate social responsibility practices, which can be used as valuable feedback to help the company improve. Second, in designing social media campaigns, companies need to be aware that likes and comments are two distinct forms of engagement that should be measured separately. The same factors can have different or even opposing effects on likes and comments. Therefore, companies should set specific goals for their social media campaigns and be cognizant of the trade-offs among different outcomes. Companies should also assess not only the sheer volume of likes or comments but the specific content that attract the likes and comments (e.g., likes of customer complaints should not be reported as a positive sign of social media marketing initiatives). Third, despite the popularity of sentiment analysis in harnessing social media data, our results suggest that there is great need and value to go beyond simple valence and to analyze the content of social media posts. Combining sentiment analysis with content analysis has the potential to reveal subtle patterns of customer behaviors to advance theory and improve practice.

Our research suggests several directions for future work. First, our empirical strategy took advantage of specific design features of Facebook business pages, i.e., user-generated posts that do not have tags to businesses and that have not been shared can *only* be seen on the business pages. Ideally, if comprehensive knowledge about the EdgeRank algorithm were available, we could explore more rigorous methods to address the endogeneity issue. In addition, future research may explore lab or field experimental setups to deal with this issue. Second, we only analyzed the textual content of a post, although some posts contained multimedia content such as photos or videos. While our primary focus in this essay is on the textual content of user-generated posts, incorporating multimedia information in the coding process is another interesting avenue for future research. Third, future research can also extend our analyses to consider small- and medium-sized businesses and nonprofit organizations, as well as taking into account the moderating roles of product and service attributes. Finally, we believe our choice of Facebook as the research context benefits the generalizability of our findings, as many other platforms often follow Facebook's design. For instance, the Like button was introduced by Facebook and later became a standard feature on many platforms (e.g., Twitter, YouTube, and Instagram). Nonetheless, future work should try to replicate our findings in other contexts.

To conclude, this is only a first step toward understanding this new form of UGC on Facebook business pages. Many companies marched into the new territory of social media marketing with limited understanding of user behaviors. Our study sheds light on the challenges and pitfalls that companies need to be aware of and prepared for. By demonstrating the distinctive nature of this new form of UGC, we hope to call for more research to understand a suite of interesting questions around it, such as the economic impact of positive and negative posts and the appropriate response or intervention strategies that companies can utilize to deal with UGC on Facebook, especially the negative ones. We believe the answers to these questions will further deepen our understanding of social media marketing and inform business practice.

Chapter 3. Dynamics of Engagement Behaviors

3.1 Introduction

The design features of a social media platform determine how it is used, by facilitating and shaping user activities and interactions (Aral et al. 2013; Sundararajan et al. 2013). For example, the iconic "Like" button that Facebook introduced in 2009 enables users to express their affection for certain content and has evolved into a ubiquitous feedback mechanism (Schöndienst et al. 2012). Besides the "Like" button, social media platforms typically also offer several other features, such as commenting, sharing, replying, and voting, to help users engage with the content and with each other. For example, the Like and Comment features are used by multiple sites, including Twitter, Instagram, LinkedIn, Google+ (in the form of "+1"), and many others. Several sites, such as YouTube, Reddit, and StackOverflow, also have a "Dislike" button, as part of a user voting system. Some product review platforms, such as Yelp, offer multidimensional rating systems, which allow users to rate a review as "funny", "useful", and "cool" to express granular emotions or opinions. Evidently, the plurality of engagement features on social media and online communities is a prevalent phenomenon.

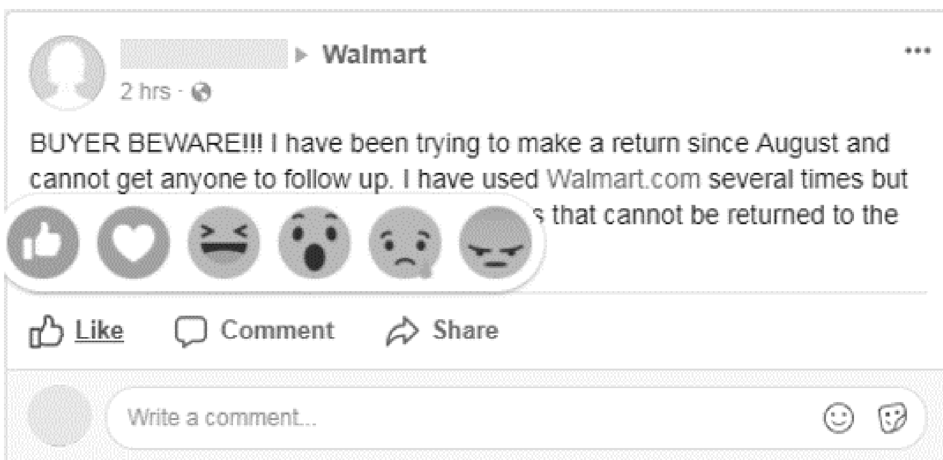
Despite the existence of multiple engagement features on social media, little is known about the relationships among these features and their usage. How do users choose which features to use to engage with content and with other users? Does having more engagement features necessarily encourage higher levels of engagement activities? Answers to these questions are unclear. On the one hand, it is possible that different engagement features are designed to provide distinctive functionalities, and users choose the features that best fit their goals and preferences. On the other hand, different engagement features may be interdependent in the sense that the usage of one feature can affect the usage of other features.

Existing literature on social media engagement behaviors (e.g., Goh et al. 2013; Rishika et al. 2013) views engagement features as given and focuses primarily on examining the antecedents or consequences of user engagement. In this essay, we aim to study the dynamic interplay among multiple engagement features, by addressing the following research questions: (1) *how does the introduction of a new engagement feature on social media platforms affect the usage of existing engagement features, and (2) how does the new feature affect the overall engagement intensity with user-generated content?*

We address our research questions using a quasi-experiment on Facebook. On February 24th, 2016, Facebook introduced the "Reactions" feature across the entire platform. In addition to Liking and Commenting, users can also engage with individual posts by clicking one of the five Reactions buttons, including *love*, *haha*, *wow*, *sad*, and *angry*. This change provides a unique opportunity to study the effect of a new engagement feature, i.e., the Reactions feature, on the usage of existing engagement features, i.e., Likes and Comments.

The context of our research is *Facebook business pages*, which are pages managed by companies and organizations on Facebook. Many companies increasingly rely on their digital presence on social media to provide marketing messages about their brands and engage their customers as fans (Goh et al. 2013; Dholakia and Durham 2010). Visitors to such business pages can view both marketer-generated (i.e., company-generated) content and user-generated content, and engage with the content through different features, including Like, Comment, or the newly introduced Reactions feature. In this essay, we focus on user engagement towards user-generated content (hereafter referred to as “user posts”), because user-generated posts outweigh marketer-generated posts on Facebook business pages in both volume and impact on consumer purchase behaviors (Goh et al. 2013). Figure 3.1 shows an example user post on Walmart’s Facebook business page and different engagement features that can be used, including Like, Comment, Share, and the newly introduced Reactions buttons (visible when hovering over the Like button).

Figure 3.1. A User Post on Facebook Business Pages and Engagement Features



We choose Facebook business pages as our empirical context for several reasons. First, user engagement on Facebook business pages is important to the focal companies, because increased engagement has been linked to growth in brand loyalty, purchase expenditures, and firm profitability (Dessart et al. 2015, Goh et al. 2013; Rishika et al. 2013). Companies have also invested heavily in social media management. Hence, studying engagement patterns on Facebook business pages has significant practical relevance. Second, while user engagement that takes place among Facebook friends is partially driven by personal relationships, which are typically unobservable to researchers, user engagement on business pages happens predominantly among strangers and is driven by measureable characteristics of the page and of the content. Such research context facilitates empirical analyses and identification.

We collected all user posts on 29 Fortune 500 companies' business pages, generated 6 months before and 6 months after the introduction of the Reactions feature. We compared user posts created before the feature change with those created after the change to identify the overall effects of Reactions on Likes and Comments for an average user post on Facebook business pages. We then separated user posts that were created *after* the feature change into two distinct groups: (1) those that have received Reactions and (2) those that have not received Reactions, and empirically examined the impact of the Reactions feature on the two groups respectively. To account for endogeneity in receiving (or not receiving) Reactions, for each of the two groups, we relied on matching methods (including propensity score matching and coarsened exact matching) to construct a matching sample of user posts *prior* to the feature change with comparable characteristics. We then examine if user posts with (or without) Reactions end up receiving more or fewer Likes and Comments than what they *would have* received before the feature change.

Our study has several key findings. Overall, the introduction of the Reactions feature increased the number of Likes but decreased the number of Comments received by an average user post on a Facebook business page. However, effect of the Reactions feature on existing engagement features is heterogeneous across different posts. Specifically, we found that the effect depends on whether the post had received any Reactions. User posts that received at least one Reaction actually ended up receiving both more Likes and more Comments than they would have received before the feature change. In contrast, user posts that were created after the feature change yet did not receive any Reactions actually ended up receiving both fewer Likes and fewer Comments than they would have received before the feature change. In other words, the introduction of the Reaction feature heightened overall engagement activities for posts that received Reactions, yet lowered overall engagement for posts that did not receive Reactions. These effects, which were already detected within 1 month after the feature change, persisted after 6 months, suggesting that the new feature led to long-term, lasting changes in overall engagement patterns.

Our work contributes new insights to the social media literature and deepens the understanding of user engagement behaviors. It is one of the first attempts to study the dynamic relationships among multiple social media engagement features. We theorize and empirically demonstrate that the introduction of a new engagement feature is not merely an additional way of user expression and interaction, but that it can also cause structural changes to engagement behaviors that may or may not have been intended by the feature designers. Our work provides practical implications for several stakeholders. For Facebook, the designer of the Reactions feature, this essay offers important empirical evidence of how the new feature is actually being used, which can inform Facebook's feature design decisions. For companies that manage their Facebook business pages, this essay uncovers the changing dynamics of user engagement behaviors, which can facilitate more effective cultivation of engagement activities on their pages.

3.2 Literature Review and Theory Development

3.2.1 Social Media Engagement Features

Likes, Comments, and the newly introduced Reactions are all instances of engagement features designed to enable users to interact with one another and express their opinions about content. In the social media and online community literature, these engagement features represent an important type of technology capability often referred to as metavoicing (Majchrzak et al. 2013; Dong et al. 2016; Nan and Lu 2014). Metavoicing allows users to engage in online conversations by reacting to other users' presence, content, and activities. By pooling individual users' opinions, metavoicing facilitates the construction of metaknowledge that may signal the quality of the content (Majchrzak et al. 2013).

Different forms of engagement features are designed to facilitate potentially different engagement intentions and goals. Taking Facebook as an example, "Like" is designed to be a one-click, lightweight feedback to demonstrate affection and enjoyment of the content (Chan 2009). In fact, in Facebook's announcement to introduce the "Like" button, it was compared to the star ratings people give on review websites (Chan 2009). In comparison, "Comment" is designed to express more substantive opinions and "longer accolades" (Gerlitz and Helmond 2011). It was thought to be analogous to textual reviews on review websites (Chan 2009). Finally, the new Reactions feature is designed to "give [users] more ways to share [their] reaction to a post in a quick and easy way" (Facebook Newsroom 2016). In other words, the Reactions feature expands the scope of emotions that can be conveyed by a one-click feedback. Beyond the affirmative emotion represented by Likes, users can now express other emotions, including more granular positive emotions (*love*, *haha*, and *wow*) as well as negative emotions (*sad* and *angry*). Meanwhile, the Reactions feature maintains the lightweight nature of engagement (similar to Like), as compared to Comment.

It is well established in the Information Systems literature that the actual usage of an artifact or technology can be different from its intended usage (e.g., Dillon and Morris 1996; Taylor and Todd 1995). Therefore, users' actual use of engagement features may deviate from what the designers have anticipated or intended. For example, Scissors et al. (2016) conducted a survey of 2,109 Facebook users to understand their perceptions of giving and getting Likes on Facebook. They found that people used the Like button for a much wider range of reasons than just expressing affection, and a Like can be perceived as a social cue that conveys feelings such as agreement, supportiveness, empathy, or simply attention.

Prior literature on social media engagement typically treats engagement features as given and focuses primarily on studying the antecedents or consequences of user engagement behaviors (e.g., Goh et al. 2013; Dessart et al. 2015; Luo et al. 2013; Rishika et al. 2013; Miller and Tucker 2013). Our work is different in that we view engagement features as dynamic, and examine how user engagement behaviors change when a new engagement feature is introduced.

Due to the potential discrepancy between intended use and actual use of the engagement features and the lack of directly relevant empirical evidence from prior literature, it is not easy to make a priori predictions regarding how the introduction of a new engagement feature would impact the existing engagement activities on social media. Therefore, we review relevant literature and propose two possible scenarios, regarding how the introduction of the Reactions feature may affect usage of Likes and Comments on Facebook business pages.

3.2.2 The Substitution Effect

There are reasons to believe that the newly introduced Reactions may substitute some Likes and Comments, so that after the introduction of the Reactions feature, the number of Likes and the number of Comments for a user post that has received Reactions may *decrease*. This happens because certain user responses previously expressed via Likes or Comments can now be directly and more appropriately expressed via Reactions. The substitution effect mainly happens at the individual user level, when a particular user's usage of Reactions affects his/her usage of Likes and Comments. For instance, a user who wants to express a strong emotion of awe towards a post would need to provide a Like and/or Comment for the post prior to the introduction of Reactions, but can now just click the "wow" button. Similarly, a user who wants to express sympathy towards a post can now simply click the "sad" button, instead of writing a comment or inappropriately clicking the Like button. In addition, the substitution effect can be explained and understood through the perspective of the *Emotion Regulation theory*.

According to the Emotion Regulation theory (Gross 1998), being able to clearly articulate one's emotion can help individuals make sense of their emotions and experiences, which subsequently leads to improved interpersonal relationships (Gross and John 2003). Therefore, having access to an engagement feature that enables easy expression of granular emotions is likely to substitute the need for other engagement features which provide lower granularity. Compared to

Likes, Reactions allow users to express more granular emotions with similar level of effort. Thus, we expect users to use Reactions, instead of Likes, when they want to convey particular emotions specified by the Reactions feature. Meanwhile, when multiple features can all fulfill the same engagement purpose, users may choose the feature that is easiest to use. Although Comments can be used to express any emotions to a high degree of clarity, writing Comments requires more effort than clicking the Reactions buttons. Thus, we expect Reactions to also substitute for Comments.

It is worth noting that the substitution effect may also be part of Facebook's intention to introduce the Reactions feature. When the Like button was introduced in 2009, it was Facebook's intention to use it to replace short, affirmative Comments such as "Awesome!" or "Congrats!" (Chan 2009; Gerlitz and Helmond 2013). Although empirical evidence is not available to test whether there was indeed substitution of Comments by Likes, it is logically plausible that users would switch to a new engagement feature when it is beneficial to their usage. Similarly, it was Facebook's intention to use the Reactions feature to give users more ways to quickly and easily respond to posts and express a wider variety of emotions, such as empathy and negative feelings.[19] As a result, users may see less need to use Likes or Comments when they want to express the types of emotions specified by the Reactions feature.

3.2.3 The Reinforcement Effect

It is also possible that Reactions may reinforce Likes and Comments, causing a post to receive more Likes and more Comments than it would have, had there been no Reactions. Unlike the substitution

[19] https://www.forbes.com/sites/kathleenchaykowski/2016/02/24/facebook-no-longer-just-has-a-like-button-thanks-to-global-launch-of-emoji-reactions

effect, the reinforcement effect can operate at both the individual user level and across multiple users, through potentially different mechanisms.

One possible driver for the reinforcement effect is the signaling mechanism under attention scarcity, which happens across different users. The abundance of user-generated content on review websites, social media platforms, and online communities implies that the content often needs to compete for users' attention (Wang et al. 2013; Shen et al. 2015; Iyer and Katona 2015). Such attention competition can be particularly intense for user posts on Facebook business pages, because the Facebook business page of a large company can receive hundreds of user posts on a daily basis, whereas visitors to the page typically spend only a few minutes during each visit (Yang et al. 2014).

Under attention scarcity, users may rely on certain quality signals to choose which content to consume and engage with. In the user-generated content literature, the intensity of engagement activities has been repeatedly used as an important signal for the popularity or quality of the content. De Vries et al. (2012) used the number of Likes and the number of Comments received by a brand post to measure the popularity of that post. Similarly, Khobzi et al. (2017) measured the counts of Likes, Comments, and Shares to gauge the extent of dissemination of a post among users. Schöndienst et al. (2012) showed that users regarded the number of Likes as a quality signal: when a post about products and services received more Likes, people perceived the quality of products and services to be superior.

On Facebook business pages, due to the abundance of user posts and the limited attention that users can spare, many users may treat the number of Likes, Comments, and Reactions as a signal of the underlying quality of the posts, and only choose to pay attention to posts that have received some engagement. Therefore, if a user post on Facebook business page has received some Reactions, we may expect it to attract even more engagement, including Likes and Comments.

Another mechanism behind the reinforcement effect is a trend-following process in participation behaviors, which also takes place across users. It has been shown in online review and online community literature that existing participations can lead to more participations. For example, Dellarocas et al. (2010) examined the creation of online reviews for motion pictures and found that people were more likely to write reviews for products that had already received many reviews. Ludford et al. (2004) studied the under-contribution problem in online communities and highlighted the importance of inspiring initial contributions, because “community activity begets activity”. Following this logic, a user post that received some Reactions may attract more Likes and Comments, simply because other users want to join the conversation and express their own opinions.

The reinforcement effect of Reactions on Comments can also happen at the individual, within-user level, driven by the complementarity between the two features. Technically speaking, the Reactions and Comments features are not mutually exclusive, in that a user can click one of the Reactions buttons *and* also write a comment. For users with a strong motivation to engage, they may choose to use both features to better express themselves. We tested this possibility in our empirical analyses.

3.2.4 A Comparison between Substitution and Reinforcement Effects

It is important to point out that the substitution and reinforcement effects, although being directionally opposite, are not mutually exclusive and may co-occur. For example, a particular user may choose to substitute Likes or Comments with Reactions based on his or her engagement goal; however, across users, Reactions may be perceived as a quality signal and attract more Likes and Comments. In this essay, our primary goal is to understand the aggregated, overall effect that Reactions may have on the use of Likes and Comments for individual posts. This is important because posts are the basic "units" of user-generated content on Facebook business pages. While different users may exhibit different engagement behaviors toward a post, it is the overall engagement (i.e., Likes, Comments, and Reactions) received by the post that indicates the impact of that post on various business-related outcomes, such as sales (Goh et al. 2013) and purchasing decisions (Rishika et al. 2013). Accordingly, we focus our analyses at the *post level*.[20]

3.2.5 Duration of Impact

An important question in the technology design and adoption literature (e.g., Bhattacherjee 2001; Kraut et al. 1998) is how long behavioral changes that are triggered by technologies last. Often the short-term usage pattern of a technology artifact can be different from its long-term usage pattern. Due to similar considerations, we are interested in studying the duration of the impact of the Reactions feature.

---

[20] As one of our empirical robustness checks, we tested the degree to which substitution effect and reinforcement effect may coexist (Section 3.5.2), but a thorough examination of the exact substitution/reinforcement process would require comprehensive data at the *individual user* level.

Immediately after the feature change, Facebook users' adoption and usage of the Reactions feature may be under the influence of the "novelty shock", i.e., the feature may be intensively used simply because it is new, and users are trying it out. The existence of such "novelty shock" is supported by a straightforward Google Trend query, which showed that the number of Google searches for the term "Facebook Reaction" spiked immediately following 02/24/2016 (i.e., the day when the Reactions feature was introduced), and then subsided after a week. This indicates that many users were actively learning about (and potentially even trying out) the new Reactions feature upon its release.

As time goes on, usage of the Reactions feature may change as the novelty effect fades, and the continued, long-term usage of the feature largely depends on user satisfaction and perceived usefulness of the feature (Bhattacherjee 2001). If the Reactions feature is perceived to be satisfactory and useful, then we can expect its usage to continue or even increase. As a result, its impact on Likes and Comments is also likely to persist over time. However, if users' experiences with the Reactions feature are unsatisfactory, then the feature may lose its attraction. Its usage will decrease, and the impact on Likes and Comments will not persist. Based on the above arguments, another question we investigate in this essay is*: Is the long-term impact of the Reactions feature on Likes and Comments similar to or different from the short-term impact?*

3.3 Empirical Context

3.3.1. Data

We collected data from the Facebook business pages of Fortune 500 companies in 6 consumer-oriented industries. We focused on Fortune 500 companies because they were early adopters of Facebook business pages, and their business pages had relatively high levels of traffic. Among all industries, we chose the 6 industries that are consumer-oriented, i.e., Airlines, Commercial Banks, Consumer Products, Food and Drug Stores, General Merchandisers, and Specialty Retailers, because the issue of user-generated content is more relevant in consumer-oriented industries than in other industries.

A total of 29 companies in these industries had active Facebook pages around the time of feature change. We collected all user-generated posts on their pages, created 6 months before and 6 months after the introduction of Reactions feature (i.e., 08/24/2015 to 08/24/2016). We constructed an unbalanced panel of 228,597 individual user posts across the 29 company pages. For each post, we collected its textual content, post type (status, video, or photo), time of creation, and the number of Likes and Comments it received. For posts created after the feature change, we also counted each type of Reactions it received. We only counted the number of Likes, Comments, or Reactions that were generated by Facebook users, rather than the focal companies. This is because (1) companies very rarely click Like or Reactions on user posts – fewer than 0.6% of user posts in our sample received any Likes or Reactions from the focal companies; and (2) companies make Comments on user posts primarily to respond to users' queries or complaints, which is outside the scope of our current work. It is worth noting that our main findings remained the same even after taking into account engagement activities from the companies.

Importantly, we identified a set of posts in our sample that were not organically generated on Facebook business pages. Specifically, 1,351 posts were created on users' own timelines with tags to the focal businesses (using the "@" sign, e.g., @WalMart). While user posts organically created on business pages are visible only to visitors to the pages,[21] posts created on users' own timelines are also exposed to the users' personal social networks, and the display of these posts are subject to Facebook's proprietary algorithm. Because the effect of Facebook's proprietary algorithm on posts' visibility is unknown to us, we removed all user posts with tags to the focal businesses from our sample.

Additionally, we do not consider the "Share" feature in the current study for several reasons. First, the Share feature is conceptually distinct from the other engagement features, because it allows users to move the content into their personal social spaces and display it to their Facebook friends (Malhotra et al. 2013). Users' motivations for sharing could be fundamentally different from their motivations to use the other types of engagement features. Second, the Share button is rarely clicked for user posts on Facebook business pages, and only 3% of the user posts in our sample received any Shares (comparatively, 35% of the user posts received at least one Like and 36% received at least one Comment). Third, even within the small set of user posts that did receive Shares, there was no significant change in the number of Shares received by a post before and after the feature change ($Mean_{before}$ = 2.04, $Mean_{after}$ = 2.14, p = 0.40). This suggests that the feature change may not have any notable impact on Shares. Fourth, for user posts that did get shared, the subsequent engagement may come from the sharer's personal Facebook network, which is

[21] Authors obtained this information by opening an actual business page on Facebook and experimenting with different ways of creating user posts.

unobservable to us. Due to the above reasons, we removed user posts that received any Shares from our sample.

3.3.2 Empirical Specifications

To identify the *overall* impact of Reactions on Likes or Comments for an average user post on Facebook business pages in the short term, we treated the introduction of Reactions feature on Facebook as a quasi-experiment (Angrist and Pischke 2008). In particular, we used the introduction of the Reactions feature on 02/24/2016 as a cutoff point, and compared posts created immediately before the change with posts created immediately after the change to assess the overall impact of the treatment, i.e., the introduction of the Reactions feature. The same methodology has been used in other studies to assess the impact of design and policy changes (e.g., Zhang and Zhu 2011; Cavusoglu et al. 2016; Kumar et al. 2017).

This empirical strategy is appropriate for our study for several reasons. First, the Reactions feature was enabled uniformly on the Facebook platform in the U.S. in a one-shot manner. Therefore, users had equal access to the Reactions feature, regardless of their geographic locations or devices. Second, although Facebook carefully planned the rollout of Reactions, there was no evidence suggesting that the rollout schedule was affected by user activities on the business pages in any way. In other words, the introduction of Reactions feature created a shock that was reasonably exogenous with respect to user activities and engagement behaviors on business pages. Third, to the best of our knowledge, user posts on Facebook business pages are not subject to Facebook's recommendation algorithm that controls the content in users' personal newsfeeds during our 12-month data collection period. Based on our observations, user posts are displayed in the "Visitor Posts" section on the business pages in reverse-chronological order. Hence, engagement with user posts is not confounded by unobserved post exposure. Fourth, there was negligible "spillover" of control units (i.e., user posts created before the change) into the treatment period. In particular, less than 0.05% (12 out of 28,995) of user posts created in the 4 weeks before the feature change continued to receive any Reactions after the feature change. We removed these 12 posts from all subsequent analyses.

Using posts published 4 weeks before and 4 weeks after the introduction of Reactions feature (N = 28,983), we estimated the following regression specification separately for Likes and Comments:

$$y_{ij} = \beta_0 + \beta_1\, After_{ij} + \beta_2\, Day_{ij} + \beta_3\, After_{ij} \times Day_{ij} + \boldsymbol{\Gamma}\, Company_j + \boldsymbol{\Phi}\, Type_{ij} + \varepsilon_{ij}$$

In the above specification, y_{ij} is the number of either Likes or Comments received by post i on the business page of company j. $After_{ij}$ is a dummy indicator of whether the post was created before or after the change; Day_{ij} represents the date on which the user post was created relative to 02/24/2016. For example, a post created on 02/23/2016 has a relative date of -1, whereas a post created on 02/25/2016 has a relative date of 1. Posts created exactly on 02/24/2016 were not included in our sample, because we do not have information about the exact time when the feature change took place on 02/24/2016. We also included the interaction term $After_{ij} \times Day_{ij}$ to control for time trends both before and after the change. To further account for unobserved heterogeneity, we included both company and post type fixed effects, represented by $Company_j$ and $Type_{ij}$ respectively. The coefficient of primary interest is β_1 (i.e., the coefficient on dummy variable $After_{ij}$), which captures the overall impact of the feature change on the dependent variable, after controlling for time trends and other factors.

Because the dependent variables are counts of Likes or Comments, the specification is estimated as Poisson regression with robust standard errors. In Table 3.1, we list some descriptive statistics regarding the distributions of Likes, Comments, and Reactions.

Table 3.1. Descriptive Statistics of Likes, Comments, and Reactions for an Average User Post (Short Term)

	Before Feature Change (4 weeks, N = 14,974)		After Feature Change (4 weeks, N = 14,009)	
	Mean	SD	Mean	SD
Likes	0.5970	1.8110	0.7274	1.8810
Comments	0.7171	1.6249	0.7306	1.5254
Reactions	NA	NA	0.0580	0.3810

After this overall effect is estimated, we further conduct sub-sample analyses using matching methods to understand the (potentially heterogeneous) effects of Reactions on Likes and Comments by splitting the sample into posts that received Reactions after the feature change and posts that did not receive any Reactions. We report more detailed analyses and all results in the next section.

3.4 Analyses and Results

3.4.1 Overall Effects on Likes and Comments

The Poisson regression results are reported in Table 3.2, showing the overall impact of the Reactions feature on Likes and Comments, respectively. Below we discuss the findings for Likes and Comments separately.

Our results show a positive treatment effect on the number of Likes ($\beta_1 = 0.4190$, $p < 0.001$).

After the introduction of the Reactions feature, an average user post received about 52% more Likes than what it would have received before the feature change.

Table 3.2. Poisson Regression Estimation Results (N = 28,983)

	DV = Likes	**DV = Comments**
After	**0.4190*** (0.0530)**	**-0.1320* (0.0529)**
Day	-0.0077** (0.0025)	0.0007 (0.0021)
After × Day	-0.0058 (0.0036)	0.0069* (0.0032)
Company fixed effects	Included	Included
Type fixed effects	Included	Included
Constant	-0.7834*** (0.1099)	-1.5804*** (0.1407)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard error in parentheses.

We conducted multiple robustness checks to verify the positive treatment effect on Likes. First, we estimated two alternative model specifications: a negative binomial regression and a linear regression with log-transformed dependent variable. Second, we repeated the analyses using time windows of 2 or 3 weeks, instead of 4 weeks, before and after the feature change. Third, we controlled for the 2nd order time trend, by adding $(Day_{ij})^2$ and $After_{ij} \times (Day_{ij})^2$ into the regression. The positive treatment effect stayed robust in all three analyses. Fourth, we conducted additional falsification tests, by artificially moving the treatment date to either 4 weeks earlier than 02/24/2016 or 4 weeks later, and estimated the "pseudo" treatment effect for each scenario. Our results showed that the "pseudo" treatment effects were not significantly positive (i.e., either insignificant or negative), which means that the positive treatment effect we found cannot merely be attributed to general time trends. Results of the above checks are summarized in Table 3.3.

In contrast, our results show a negative treatment effect on the number of Comments (β_1 = -0.1320, $p < 0.05$). After the introduction of the Reactions feature, an average user post received about 12.4% fewer Comments than what it would have received before the feature change. We also conducted robustness checks on this effect, including: (a) estimating alternative models of a negative binomial regression and a linear regression with log-transformed dependent variable; (b) using alternative time windows of 2 or 3 weeks instead of 4 weeks; (c) controlling for the 2nd order time trend; and (d) running falsification tests with pseudo treatment dates. The results are shown in Table 3.4. In columns 1-5 of Table 3.4, the negative treatment effect stayed directionally consistent. In columns 6 and 7, the estimated "pseudo" treatment effects were positive and insignificant, indicating that the negative treatment effect we found cannot be attributed merely to general time trends.

Table 3.3. Robustness Checks for the Effect on Likes

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
After	**0.2860*****	**0.0934*****	**0.2323****	**0.4831*****	**0.2835****	**-0.1068^{+}**	**0.0356**
	(0.0480)	**(0.0117)**	**(0.0821)**	**(0.0674)**	**(0.0912)**	**(0.0619)**	**(0.0756)**
Day	-0.0055**	-0.0013**	0.0017	-0.0182***	-0.0129	-0.0005	-0.0161***
	(0.0021)	(0.0005)	(0.0083)	(0.0051)	(0.0118)	(0.0021)	(0.0028)
After × Day	-0.0043	-0.0026***	0.0036	0.0067	0.0337*	-0.0047	0.0398***
	(0.0031)	(0.0007)	(0.0102)	(0.0062)	(0.0162)	(0.0035)	(0.0042)
Company fixed effects	Included	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included	Included
Constant	-0.8119***	0.2407***	-0.8131***	-0.8658***	-0.8191***	-0.2089^{+}	-0.5645***
	(0.1088)	(0.0272)	(0.1308)	(0.1230)	(0.1251)	(0.1206)	(0.1239)
N	28,983	28,983	15,241	22,127	28,983	31,983	26,520

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. Column definitions: (1) negative binomial regression; (2) OLS with log-transformed DV; (3) alternative time window, 2 weeks; (4) alternative time window, 3 weeks; (5) control for 2nd order time trend; (6) pseudo treatment effect on 01/27/2016; (7) pseudo treatment effect on 03/23/2016.

3.4.2 Sub-Sample Analyses

The above regression analysis provides strong evidence of an overall positive treatment effect of

Reactions on Likes and an overall negative treatment effect on Comments, which appears to suggest that the newly introduced Reactions might be reinforcing Likes but substituting for Comments. However, it is important to note that not all user posts that were created after the feature change received Reactions. We further analyzed the effect of the Reactions feature for the sub-sample of posts that actually received Reactions.

Table 3.4. Robustness Checks for the Effect on Comments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
After	**-0.0896^{+}**	**-0.0313***	**-0.1486***	**-0.1156^{+}**	**-0.1322**	**0.0196**	**0.0001**
	(0.0614)	**(0.0130)**	**(0.0758)**	**(0.0616)**	**(0.0824)**	**(0.0481)**	**(0.0524)**
Day	-0.0008	0.0003	0.0110^{+}	0.0021	0.0090	-0.0052**	0.0086***
	(0.0021)	(0.0005)	(0.0065)	(0.0037)	(0.0097)	(0.0018)	(0.0025)
After × Day	0.0083**	0.0026**	-0.0064	0.0021	-0.0099	0.0061*	-0.0151***
	(0.0031)	(0.0008)	(0.0090)	(0.0050)	(0.0134)	(0.0029)	(0.0035)
Company fixed effects	Included	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included	Included
Constant	-1.5933***	0.0668**	-1.5566***	-1.6878***	-1.5348***	-1.4187***	-1.1771***
	(0.1423)	(0.0228)	(0.1781)	(0.1592)	(0.1510)	(0.1395)	(0.1663)
N	28,983	28,983	15,241	22,127	28,983	31,983	26,520

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. Column definitions: (1) negative binomial regression; (2) OLS with log-transformed DV; (3) alternative time window, 2 weeks; (4) alternative time window, 3 weeks; (5) control for 2nd order time trend; (6) pseudo treatment effect on 01/27/2015; (7) pseudo treatment effect on 03/23/2016.

If Reactions reinforced Likes, then we should expect posts that received Reactions to have more Likes than what they would have received before the feature change. Similarly, if Reactions substituted Comments, then we should expect posts that received Reactions to have fewer Comments than what they would have received before the feature change. To test these claims, we partitioned the sample of user posts created within 4 weeks after feature change into two non-overlapping subsamples: (1) 591 posts that received at least one of the five new Reactions and (2)

13,418 posts that did not receive any Reactions. We estimated the treatment effects for each of the two subsamples, respectively.

Directly comparing posts that received Reactions with pre-treatment posts suffers from the endogeneity issue, because key characteristics of a post, such as its content quality or popularity, may affect both the number of Likes/Comments and the number of Reactions it receives. Therefore, we used propensity score matching (Rosenbaum and Rubin 1983) to construct a proper control group. Each post that received Reactions was matched with the nearest counterfactual among pre-treatment posts, based on 7 post-level observables described as follows:

- The company that owns the page where the post appeared;
- Post type, one of status, video, or photo;
- Post length, measured as word count;
- Post sentiment variables: (1) percentage of positive words in the post and (2) percentage of negative words in the post, obtained by analyzing the textual content of the user posts using the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al. 2007);
- Post contextual variables: (1) the number of other user-generated posts created within 24 hours before and 24 hours after the creation of the focal post and (2) the number of marketer-generated posts created within 24 hours before and 24 hours after the creation of the focal post. We use these two variables to measure the number of other posts that might compete for users' attention with the focal post on the business page.

The propensity score was estimated with a logistic regression, and the matching was done using the nearest-neighbor approach. The 591 posts that received Reactions were matched with 591 pre-treatment counterfactual posts. We then estimated the treatment effect by comparing the two matched samples with a t-test.

We found that, while a post in the pre-treatment matched sample received 1.0795 Likes on average, a post with Reactions received 2.7919 Likes on average (t = 6.82, p < 0.001), indicating a strongly positive treatment effect. Interestingly, we also found a *positive* treatment effect on the number of Comments. While a post in the pre-treatment matched sample received 0.9509 Comments on average, a post with Reactions received 1.8917 Comments on average (t = 6.56, p < 0.001). Therefore, posts with at least one Reaction actually received more Likes and more Comments than they would have before the feature change, supporting a reinforcement effect (rather than a substitution effect) for both Likes and Comments.

To assess the quality of our propensity score matching process, we estimated the Rosenbaum Bound (Rosenbaum 2002) as a measure of the sensitivity of our findings with respect to unobserved selection variables. The Rosenbaum Bound of a particular treatment effect estimation has a critical value, which numerically represents how likely the estimated effects can be invalidated by unobserved variables that affect the odds of selection into the treatment group. For analysis of Likes, the matching process has a critical Rosenbaum Bound of 3.0, suggesting that unobserved variables would have to alter the odds of selection into treatment group by 200% in order to invalidate the treatment effect. For analysis of Comments, the critical value of Rosenbaum Bound was 3.6, suggesting that unobserved variables would have to alter the odds of selection into treatment group by 260% in order to invalidate the treatment effect. Both values indicate our findings are highly unlikely to be nullified by unobserved selection variables.

We repeated the matching analyses with an alternative matching approach, known as the Coarsened Exact Matching (Iacus et al. 2012). Instead of matching treatment units with control units based on propensity scores, this approach seeks *exact* matches between treatment and control units, based on discretized ("coarsened") matching variables. It has been shown to outperform propensity score matching in producing a more balanced matched sample (Iacus et al. 2012; King and Nielsen 2016). We obtained qualitatively similar results for both Likes and Comments. Specifically, 433 (out of 591) posts that received Reactions were matched with 7,809 pre-treatment counterfactual posts. Comparing these two matched samples, posts with Reactions received significantly more Likes ($Mean_{before}$ = 0.5536, $Mean_{after}$ = 2.2171, $p < 0.001$), as well as more Comments ($Mean_{before}$ = 0.8191, $Mean_{after}$ = 2.1663, $p < 0.001$). Coarsened exact matching also enables estimation of average treatment effect on the treated (ATT). ATT is estimated by first calculating the treatment effects between each treated unit and its matched control units, and then taking average of those treatment effects over all treated units. Using our two matched sample, we estimated positive ATT for both Likes and Comments (ATT_{Likes} = 1.1308, $p < 0.001$; $ATT_{Comments}$ = 1.2084, $p < 0.001$). These additional results from coarsened exact matching further demonstrated the robustness of our findings.

In addition, we tested the robustness of the above findings against the design complementarity of Reactions and Comments. For a given post, while a user *cannot* click both one of the Reactions buttons and the Like button, he/she *can* both click a Reactions button and leave a Comment. In other words, the observed increase in the number of Comments for user posts with Reactions might be attributed to the fact that many users used the Reactions feature *and* the Comment feature on the same posts together. To test this possibility, we re-ran the matching analysis for Comments, but excluded all the Comments that were made by users who had also clicked the Reactions on the same posts. We obtained similar findings, that is, posts that received Reactions still ended up receiving more Comments than pre-treatment matched sample ($Mean_{before}$ = 0.9509, $Mean_{after}$ = 1.7360, $p < 0.001$). This suggests that the increase in Comments was not due to the simultaneous use of both Reactions and Comments by the same users.

In another set of analyses, we compared the 13,418 posts that did *not* receive any Reactions with matched pre-treatment posts, obtained with coarsened exact matching. We used coarsened exact matching instead of propensity score matching in this case, because the latter failed to reach a reasonable Rosenbaum sensitivity bound, indicating unstable matching results. Specifically, 10,627 (out of 13,418) posts that did not receive Reactions were matched with 11,740 pre-treatment counterfactual posts. We found that, while a post in the pre-treatment matched sample received 0.5577 Likes on average, a post without any Reaction received only 0.4882 Likes on average (t = 3.57, $p < 0.01$), indicating a negative treatment effect. We also found a negative treatment effect on Comments. While a post in the pre-treatment matched sample received 0.7575 Comments on average, a post without any Reaction received only 0.7047 Comments on average (t = 2.52, $p <$ 0.01). Consistent with the t-test results, using the two matched samples, we estimated negative ATT for both Likes and Comments (ATT_{Likes} = -0.0362, $p < 0.05$; $ATT_{Comments}$ = -0.0479, $p < 0.05$).

To further establish the validity of the above findings, we replicated the sub-sample analyses with regression-based approaches. Specifically, we repeated the sub-sample analyses, by *controlling for* all 7 post-level characteristics, instead of *matching on* them, using the following regression specification:

$$y_{ij} = \beta_0 + \beta_1 After_{ij} + \beta_2 Length_{ij} + \beta_3 Positive_{ij} + \beta_4 Negative_{ij} + \beta_5 UGC_{ij} + \beta_6 MGC_{ij}$$
$$+ \boldsymbol{\Gamma} Company_j + \boldsymbol{\Phi} Type_{ij} + \varepsilon_{ij}$$

Here, $Length_{ij}$ represents the word count of the post. $Positive_{ij}$ and $Negative_{ij}$ represent the percentage of positive/negative words in the post. UGC_{ij} and MGC_{ij} represent the number of other user-generated/marketer-generated posts created within 24 hours before and 24 hours after the creation of the focal post. $Company_j$ and $Type_{ij}$ represent company and post type fixed effects, respectively. The results were directionally consistent with matching results, as summarized in Table 3.5.

Overall, our sub-sample analyses suggest that the introduction of Reactions feature induced heterogeneous effects on two different sub-samples of user posts on Facebook business pages. Posts that received Reactions ended up also receiving more Likes and more Comments than what they would have received before the feature change. In contrast, posts that were created after the change and that did not receive any Reactions ended up receiving even fewer Likes and fewer Comments than they would have received before the feature change. Thus, the introduction of the Reactions feature reinforced engagement for user posts that had received Reactions yet cannibalized engagement from the posts that did not receive any Reactions.

Table 3.5. Sub-Sample Regression Estimation Results

	DV = Likes		**DV = Comments**	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**1.0304*** (0.0952)**	**-0.0299 (0.0303)**	**0.8188*** (0.0685)**	**-0.0721** (0.0259)**
Length	0.0007*** (0.0001)	0.0005*** (0.0001)	0.0005*** (0.0001)	0.0004*** (0.0001)
Positive	0.0084*** (0.0024)	0.0044* (0.0020)	-0.0401*** (0.0038)	-0.0377*** (0.0029)
Negative	0.0078^{+} (0.0040)	0.0128*** (0.0019)	0.0138* (0.0069)	0.0090* (0.0041)
UGC	0.0013* (0.0006)	0.0020*** (0.0002)	-0.0034*** (0.0006)	-0.0028*** (0.0003)
MGC	0.0523^{+} (0.0280)	-0.0090 (0.0155)	0.0129 (0.0214)	0.0186 (0.0135)
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	-1.1428*** (0.1794)	-0.9411*** (0.1104)	-1.1415*** (0.2137)	-1.2900*** (0.1468)
N	15,565	28,392	15,565	28,392

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

It is important to note that the results from sub-sample analyses do not contradict the overall effects we showed in Section 3.4.1. The overall increase of Likes for an average user post was driven primarily by the strong positive treatment effect on posts that received Reactions. In contrast, the overall decrease in Comments for an average user post was driven by the negative treatment effect on a large number of posts created after the change that did not receive any Reactions. In other words, the overall effects on Likes and Comments depend on the interplay of the two opposite forces on the two sub-samples. For posts that received Reactions, the reinforcement effect on Likes is stronger in magnitude than the reinforcement effect on Comments, which causes the manifested overall effects of Reactions to be positive for Likes but negative for Comments.

3.4.3 Analyses of Impact Duration

So far we have only considered the short-term immediate effects of Reactions on Likes and Comments, with a focus on identifying the *direction* of effects. In this section, we analyze the *duration* of effects. Our analyses consist of two steps. In the first step, we still took the sample of user posts created 4 weeks before and 4 weeks after the feature change, and checked the effects of the Reactions feature against two possible confounding scenarios: (1) the existence of a “novelty shock”, such that the observed effects of the Reaction feature on Likes and Comments were purely driven by the short-term usage of this new feature immediately after its introduction; and (2) the existence of an “anticipation effect”, such that users’ engagement behaviors might be structurally different as the rollout of Reactions became close, in anticipation of the feature change. In the second step, we expanded the time window of our analyses to include all posts created 6 months before and 6 months after the feature change. We also repeated the sub-sample analyses to check heterogeneity in the long-term effects of the Reactions feature for posts that received vs. did not receive Reactions.

To check against the “novelty shock”, we re-estimated the sub-sample regression (discussed in Section 3.4.2) using all 4 weeks before the feature change, but only the 3rd and 4th week after the change (i.e., removing the first two weeks immediately following the change). The results are reported in Table 3.6. We continued to find significant positive effects on Likes and

Comments for posts that received Reactions, as well as significant negative effects on Likes and Comments for posts that did not receive Reactions. Therefore, our findings cannot be simply attributed to a short-term "novelty shock" by the new Reactions feature.

Table 3.6. Sub-Sample Regression Results, Removing First 2 Weeks after Feature Change

	DV = Likes		**DV = Comments**	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**0.9599*** (0.1297)**	**-0.1609*** (0.0396)**	**0.7849*** (0.0890)**	**-0.0719* (0.0319)**
Length	0.0007*** (0.0001)	0.0005*** (0.0001)	0.0005*** (0.0001)	0.0004*** (0.0001)
Positive	0.0070** (0.0025)	0.0067*** (0.0021)	-0.0410*** (0.0040)	-0.0411*** (0.0034)
Negative	0.0124*** (0.0035)	0.0155*** (0.0028)	0.0142* (0.0071)	0.0125* (0.0057)
UGC	-0.0009 (0.0012)	0.0013 (0.0009)	-0.0031*** (0.0008)	-0.0021** (0.0007)
MGC	0.0585* (0.0290)	-0.0048 (0.0190)	0.0243 (0.0219)	0.0189 (0.0159)
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	-1.0122*** (0.2169)	-1.0885*** (0.1826)	-1.2124*** (0.2279)	-1.3643*** (0.1966)
N	15,264	21,614	15,264	21,614

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

Next, to check against the "anticipation effect", we repeated the above analyses using the 3rd and 4th week both before and after the feature change (i.e., removing the 4 weeks immediately adjacent to the feature change). Our main results again remained directionally consistent, ruling out the "anticipation effect". The results are reported in Table 3.7.

In the second step, we expanded the time window to include user posts created 6 months before and 6 months after the feature change. Table 3.8 shows descriptive statistics regarding the distributions of the Likes, Comments, and Reactions of all posts. According to Table 3.8, we can see that an average post received more Reactions in 6 months after the feature change than what they received in 4 weeks after the change. Both the number and the percentage of posts that received at least one Reaction were higher in the 6 month window (7,529 out of 98,321, or 7.66%) than in the 4 weeks window (591 out of 14,009, or 4.22%). This provides descriptive evidence that the use of the Reactions feature increased in the long term.

Table 3.7. Sub-Sample Regression Results, Removing Immediate 4 Weeks around Feature Change

	DV = Likes		**DV = Comments**	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**0.9470*** **	**-0.2080*** **	**0.8217*** **	**-0.0514**
	(0.1450)	**(0.0467)**	**(0.0973)**	**(0.0377)**
Length	0.0007***	0.0004***	0.0006***	0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Positive	0.00751*	0.0072**	-0.0334***	-0.0362***
	(0.0033)	(0.0026)	(0.0055)	(0.0041)
Negative	0.0128**	0.0164***	0.0171^{+}	0.0140^{+}
	(0.0044)	(0.0033)	(0.0098)	(0.0072)
UGC	-0.0001	0.0030*	-0.0028*	-0.0012
	(0.0019)	(0.0012)	(0.0013)	(0.0009)
MGC	0.0753^{+}	-0.0147	0.0108	0.0102
	(0.0397)	(0.0236)	(0.0289)	(0.0180)
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	-1.0028***	-1.0781***	-1.1133***	-1.3613***
	(0.2959)	(0.2241)	(0.3030)	(0.2400)
N	8,123	14,473	8,123	14,473

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

Next, we conducted regression analyses to estimate the long-term effects. We controlled for the same 7 post-level characteristics that were used for matching in the short-term analyses and used the following model specification:

$$y_{ij} = \beta_0 + \beta_1\, After_{ij} + \beta_2\, Length_{ij} + \beta_3\, Positive_{ij} + \beta_4\, Negative_{ij} + \beta_5\, UGC_{ij} + \beta_6\, MGC_{ij}$$

$$+ \boldsymbol{\Psi}\, Month_{ij} + \boldsymbol{\Gamma}\, Company_j + \boldsymbol{\Phi}\, Type_{ij} + \varepsilon_{ij}$$

Table 3.8. Descriptive Statistics of Likes, Comments, and Reactions for an Average User Post (Long-Term)

	Before Feature Change (6 months, N = 128,398)		After Feature Change (6 months, N = 98,321)	
	Mean	SD	Mean	SD
Likes	0.8505	2.2761	0.7595	2.0401
Comments	0.7626	1.6215	0.7951	1.7765
Reactions	NA	NA	0.1136	0.5255

In addition to the post-level controls, we also included month fixed effects, represented by $Month_{ij}$, to control for unobserved time heterogeneity at the month level. Similar to our short-term analyses, β_1 captures the impact of Reactions on the dependent variable. A similar approach has been adopted by Cavusoglu et al. (2016) in understanding the long-term impact of a particular policy change on Facebook. The specification was estimated as a Poisson regression with robust standard errors. The regression results are reported in Table 3.9.

Several things are worth noting. First, for posts that received at least one Reaction, we found a strong positive effect on Likes (β_1 = 0.1924, p < 0.001), and a strong positive effect on Comments (β_1 = 1.0833, p < 0.001). User posts that received Reactions ended up receiving more Likes and Comments than what they would have received before the feature change. This is consistent with our findings in the short-term analyses, indicating that the impact of the Reactions feature is likely to persist in the long term.

Table 3.9. Long-Term Effects Estimation Results

	DV = Likes		DV = Comments	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**0.1924*** (0.0452)**	**-0.7464*** (0.0252)**	**1.0833*** (0.0309)**	**-0.1798*** (0.0226)**
Length	0.0003** (0.0001)	0.0003*** (0.0000)	0.0004*** (0.0001)	0.0004*** (0.0001)
Positive	0.0088*** (0.0009)	0.0073*** (0.0007)	-0.0299*** (0.0014)	-0.0325*** (0.0011)
Negative	0.0118*** (0.0011)	0.0146*** (0.0009)	0.0056*** (0.0013)	0.0045*** (0.0010)
UGC	0.0004*** (0.0000)	0.0005*** (0.0000)	-0.0002*** (0.0000)	-0.0001*** (0.0000)
MGC	-0.0620*** (0.0060)	-0.0671*** (0.0048)	-0.0009 (0.0047)	-0.0030 (0.0039)
Month fixed effects	Included	Included	Included	Included
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	0.0268 (0.0569)	0.0038 (0.0501)	-1.1376*** (0.0609)	-1.3630*** (0.0541)
N	135,927	219,190	135,927	219,190

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

For posts that did not receive any Reactions, we found a significant negative effect on Likes (β_1 = -0.7464, $p < 0.001$), and a significant negative effect on Comments (β_1 = -0.1798, $p < 0.001$). In other words, a post created after the change that did not receive any Reactions ended up receiving fewer Likes and fewer Comments than it would have before the change. Again, the long-term effects on Likes and Comments are consistent with the short-term effects, suggesting a persistent impact of the Reactions feature over time.

We repeated the above analyses using two different matching methods and obtained qualitatively similar results. The results from both propensity score matching and coarsened exact matching are summarized in Table 3.10. In addition, we calculated the Rosenbaum sensitivity bounds of the propensity score matching processes. For the matching process regarding posts with Reactions, the critical Rosenbaum bound is 2.5 for Likes and 3.2 for Comments. For the matching process regarding posts without Reactions, the critical Rosenbaum bound is 1.2 for Likes and 1.1 for Comments. While the critical bounds for posts without Reactions are relatively low, the associated findings can be replicated by coarsened exact matching, indicating the overall robustness of our findings.

Table 3.10. Long-Term Matching Analyses Results

		Posts with Reactions		Posts without Reactions	
		Likes	Comments	Likes	Comments
Propensity Score Matching	Average before change	1.4428	0.9420	0.8125	0.7165
	Average after change	2.6419	2.3935	0.6034	0.6625
	t-test	***	***	***	***
Coarsened Exact Matching	Average before change	0.5659	0.9419	0.5331	0.8020
	Average after change	2.2599	2.0837	0.4723	0.7264
	t-test	***	***	***	***
	ATT estimate	1.5073***	1.6228***	-0.1571***	-0.0385***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.5 Additional Robustness Checks and Analysis

The above analyses have offered ample quantitative evidence that getting Reactions leads a user post to receive more Likes and Comments than it would have received prior to the feature change, and such reinforcement relationship manifests both in the short term and in the long run. In this section, we conducted three additional sets of robustness checks to further validate our findings and provide some additional insights. In the first set of robustness checks, we re-ran the analyses and explicitly accounted for the correlation between getting Likes and getting Comments. In the second set of robustness checks, we used exploratory content analyses to gauge the extent to which the reinforcement effect of Reactions may co-occur with the potential substitution effect. In the third set of robustness checks, we employed quantile regressions to estimate the effects of the Reactions feature on Likes and Comments in both short term and long term.

3.5.1 Accounting for the Interplay between Likes and Comments

In our main analyses, we have treated the number of Likes and Comments received by a user post as two independent measures of engagement activities. However, they can be correlated with each other, as user posts that received Likes might be more likely to attract subsequent Comments and vice versa. To account for such correlation, we conducted the robustness checks where we added the number of Likes as an independent variable into regressions on Comments, and similarly, added the number of Comments into regressions on Likes. This was done for both short-term and long-term analyses, and the results are reported in Tables 3.11 and 3.12. Our results remained largely unchanged. In particular, although there were two instances where the coefficient on *After* became statistically insignificant, there was no change in the direction of the effects. Meanwhile, coefficients on *Likes* and *Comments* are positive and significant in their respective regressions, indicating that there is indeed positive association between Likes and Comments for user posts on Facebook business pages.

3.5.2 Exploratory Content Analyses

Although our analyses showed an overall reinforcement effect of the Reactions feature on Likes and Comments for posts that received Reactions, they do not entirely exclude the likelihood of a substitution effect. As we have pointed out earlier (Section 3.2.4), it is possible that the

Table 3.11. Short-Term Results Accounting for Correlation between Likes and Comments

	DV = Likes		DV = Comments	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**0.9447*** (0.0938)**	**-0.0143 (0.0302)**	**0.7291*** (0.0709)**	**-0.0633* (0.0258)**
Length	0.0007*** (0.0001)	0.0004*** (0.0001)	0.0005*** (0.0001)	0.0004*** (0.0001)
Positive	0.0103*** (0.0024)	0.0061** (0.0020)	-0.0423*** (0.0040)	-0.0377*** (0.0030)
Negative	0.0049 (0.0045)	0.0121*** (0.0020)	0.0144* (0.0069)	0.0088* (0.0041)
UGC	0.0016*** (0.0005)	0.0022*** (0.0002)	-0.0034*** (0.0006)	-0.0029*** (0.0003)
MGC	0.0514^{+} (0.0272)	-0.0127 (0.0154)	0.0065 (0.0215)	0.0215 (0.0135)
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	-1.1829*** (0.1768)	-0.9694*** (0.1099)	-1.1512*** (0.2121)	-1.3073*** (0.1463)
Likes	-	-	0.0457*** (0.0062)	0.0572*** (0.0086)
Comments	0.0770*** (0.0093)	0.0831*** (0.0088)	-	-
N	15,264	21,614	15,264	21,614

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

reinforcement effect and the substitution effect co-exist and the two jointly determine the overall engagement for a user post. For example, for a given user post, some Likes and Comments may have been substituted by Reactions, but even more Likes and Comments may be generated due to the reinforcement effect of Reactions, resulting in an overall increase in Likes and Comments. In this section, we conducted an exploratory content analyses of Comments to further understand the interplay between the reinforcement and substitution effects. In particular, we attempt to offer some preliminary understanding of the magnitude and scale of the potential *substitution* effect at the individual post level.

Table 3.12. Long-Term Results Accounting for Correlation between Likes and Comments

	DV = Likes		**DV = Comments**	
	Posts with Reactions	Posts without Reactions	Posts with Reactions	Posts without Reactions
After	**0.0515** **(0.0451)**	**-0.7164***** **(0.0256)**	**1.0658***** **(0.0307)**	**-0.1417***** **(0.0224)**
Length	0.0003*** (0.0000)	0.0003*** (0.0000)	0.0004*** (0.0001)	0.0004*** (0.0001)
Positive	0.0104*** (0.0010)	0.0091*** (0.0007)	-0.0320*** (0.0016)	-0.0346*** (0.0013)
Negative	0.0113*** (0.0011)	0.0143*** (0.0009)	0.0057*** (0.0013)	0.0037*** (0.0010)
UGC	0.0004*** (0.0000)	0.0005*** (0.0000)	-0.0002*** (0.0000)	-0.0001*** (0.0000)
MGC	-0.0586*** (0.0059)	-0.0655*** (0.0049)	0.0017 (0.0047)	0.0000 (0.0039)
Month fixed effects	Included	Included	Included	Included
Company fixed effects	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included
Constant	0.0216 (0.0564)	0.0010 (0.0498)	-1.1789*** (0.0603)	-1.4165*** (0.0535)
Likes	-	-	0.0412*** (0.0039)	0.0592*** (0.0024)
Comments	0.0894*** (0.0042)	0.1009*** (0.0051)	-	-
N	135,927	219,190	135,927	219,190

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

If any substitution effect does occur on a substantial scale, then we should observe systematic differences in the Comments associated with posts before the feature change and the Comments associated with posts after the feature change that received Reactions. Specifically, if the Reactions feature was used as intended, they should have substituted short comments such as "Love it!", "so sad", or "I feel angry!" As a result, the remaining comments should be longer on average, and should contain fewer mentions of "love", "haha", "wow", "sad", and "angry" as well as the synonyms and variations of these words, than what the comments would have been before the introduction of Reactions.

We took the 5,924 user posts created within the 6 months after feature change that had at least one Reaction *and* at least one Comment, then created a matched sample among posts before the feature change. The matching process was the same as discussed before. Then, we compared the Comments associated with user posts in the two matched samples on the following attributes: (1) average length of Comments; (2) average counts of each of the five words: “love”, “haha”, “wow”, “sad”, and “angry”, as well as their synonyms and simple variations among Comments. More specifically, for each of the above five words, we counted direct mentions of it, mentions of its synonyms (e.g., “affection” was counted as mentions of “love”), and mentions of its simple variations (i.e., any words that have the focal word as a sub-string, e.g., “hahaha” was counted as mentions of “haha”). We provide the list of synonyms considered in Table 3.13. The results of the analysis are summarized in Table 3.14.

Table 3.13. Synonyms Considered in Exploratory Content Analyses of Comments

Keyword	Synonyms
love	affection, attachment, devotedness, devotion, fondness, passion, beloved, darling, dear, flame, hon, honey, sweetheart, sweet, sweetie, fancy, favor, like, liking
haha	humor, laugh, trick, play, antic, gag, farce
wow	wow, amuse, delight, charm, entertain, awe
sad	bad, blue, brokenhearted, crestfallen, dejected, depressed, despondent, disconsolate, doleful, down, downcast, downhearted, droopy, forlorn, gloomy, glum, heartbroken, heartsick, heartsore, heavyhearted, inconsolable, joyless, melancholic, melancholy, miserable, mournful, saddened, sorrowful, sorry, unhappy, woebegone, woeful, wretched
angry	angered, apoplectic, ballistic, choleric, enraged, foaming, fuming, furious, hopping, incensed, indignant, inflamed, infuriate, infuriated, irate, ireful, livid, mad, outraged, rabid, rankled, riled, riley, roiled, sore, steaming, ticked, wrathful, wroth

Note. Synonyms of words “love”, “sad”, and “angry” were obtained by looking up Merriam-Webster thesaurus; synonyms of words “haha” and “wow” were obtained from Thesaurus.com.

Table 3.14. Exploratory Content Analyses of Comments

	Propensity Score Matching		
	Average before change	Average after change	t-test
Average Comment Length	35.10	29.77	***
Average counts of *love* and synonyms and variations	0.1979	0.1675	***
Average counts of *haha* and synonyms and variations	0.0271	0.0244	n.s.
Average counts of *wow* and synonyms and variations	0.0129	0.0195	***
Average counts of *sad* and synonyms and variations	0.1356	0.1370	n.s.
Average counts of *angry* and synonyms and variations	0.0571	0.0466	*

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. not significant

According to Table 3.14, there were systematic differences in Comments of pre-treatment user posts and comparable post-treatment user posts that received Reactions. However, the differences were not all consistent with our expectations. On the one hand, comparing the two matched samples, Comments after the feature change contained fewer mentions of words "love" and "angry" (including their synonyms and variations), suggesting that the new "love" and "angry" buttons may have substituted some Comments. On the other hand, contrary to our expectations, Comments after the feature were *shorter* on average, and contained more mentions of the word "wow" (including its synonyms and variations). In summary, while we found some initial evidence of Reactions substituting certain Comments, the feature did not systematically substitute for short Comments on a significant scale, and not all of its five buttons were significantly substituting Comments with overlapping emotional words. This set of exploratory content analyses indicates that, at the individual post level, even if there is a co-existence of reinforcement and substitution effects of the Reactions feature, the substitution effect is relatively weak.

3.5.3 Quantile Regressions

In this sub-section, we repeated our analyses with quantile regressions. Compared with the generalized OLS method (e.g., the Poisson regressions we used in the main analyses), which estimates the conditional *mean* of dependent variables given values of independent variables, the quantile regression method estimates the conditional *quantiles* of dependent variables given independent variables. Quantile regression has two major advantages over OLS, both of which are relevant in our context. First, it is less susceptible to outliers in dependent variables, resulting in more robust coefficient estimates even when a small number of posts received a disproportionally high amount of engagement (Gao et al. 2015). Second, whereas Poisson regression estimates the *average* treatment effect among all user posts, the quantile regression allows us to examine whether the treatment effect stays robust across user posts with different levels of "popularity", as indicated by the amount engagement they received (Oestreicher-Singer and Sundararajan 2012). If the estimated treatment effects across multiple quantiles are consistent, it will provide additional validation of our findings and is helpful in obtaining a more comprehensive view of the treatment effect.

We ran quantile regressions for count dependent variables (Geraci 2016) on both short-term and long-term data, separately for user posts with and without Reactions. Because the distributions of Likes and Comments across user posts are extremely skewed (i.e., following the "long-tail" shape), we estimated the regressions at 90%, 95%, and 99% quantiles. The short-term and long-term results are included in Tables 3.15 and 3.16, respectively.

The short-term regression results in Table 3.15 show that, for user posts that received Reactions, the treatment effects on Likes and Comments are significantly positive across all three quantiles.

Table 3.15. Short-Term Quantile Regression Estimation Results

	DV = Likes					
	Posts with Reactions			Posts without Reactions		
	90%	95%	99%	90%	95%	99%
After	**1.1065*** (0.0953)**	**1.2221*** (0.1736)**	**1.2966*** (0.2742)**	**-0.0112 (0.0246)**	**-0.0145 (0.0285)**	**-0.0317 (0.0506)**
Length	0.0018*** (0.0002)	0.0014*** (0.0002)	0.0012^{+} (0.0006)	0.0016*** (0.0002)	0.0016*** (0.0001)	0.0014** (0.0004)
Positive	0.0095** (0.0036)	0.0097 (0.0062)	0.0044 (0.0046)	0.0074** (0.0028)	0.0060** (0.0022)	0.0117^{+} (0.0064)
Negative	0.0297*** (0.0046)	0.0183** (0.0066)	0.0141^{+} (0.0084)	0.0314 (0.0038)	0.0263 (0.0037)	0.0249*** (0.0053)
UGC	0.0026*** (0.0005)	0.0016* (0.0007)	0.0001 (0.0017)	0.0014 (0.0002)	0.0012 (0.0002)	0.0003 (0.0004)
MGC	-0.0047 (0.0168)	-0.0104 (0.0204)	0.0389 (0.0396)	-0.0315** (0.0119)	-0.0222* (0.0113)	-0.0053 (0.0272)
Company fixed effects	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included
Constant	-0.4298** (0.1353)	0.2496 (0.1864)	0.9609** (0.3150)	-0.1036 (0.0975)	0.4064*** (0.1107)	1.1537*** (0.2137)
	DV = Comments					
	Posts with Reactions			Posts without Reactions		
	90%	95%	99%	90%	95%	99%
After	**0.9125*** (0.1158)**	**0.8823*** (0.0913)**	**0.5294*** (0.1421)**	**-0.0295 (0.0264)**	**-0.0039 (0.0317)**	**-0.0576 (0.0496)**
Length	0.0014*** (0.0003)	0.0017*** (0.0003)	0.0014** (0.0005)	0.0018*** (0.0002)	0.0019*** (0.0002)	0.0018** (0.0006)
Positive	-0.0474*** (0.0046)	-0.0406*** (0.0042)	-0.0307*** (0.0064)	-0.0398*** (0.0039)	-0.0301*** (0.0044)	- 0.0188*** (0.0048)
Negative	0.0150* (0.0053)	0.0099^{+} (0.0052)	-0.0053 (0.0130)	0.0121** (0.0040)	0.0134* (0.0053)	0.0132 (0.0099)
UGC	-0.0028*** (0.0008)	-0.0024*** (0.0006)	-0.0017 (0.0011)	-0.0023*** (0.0003)	-0.0022*** (0.0003)	-0.0012** (0.0004)
MGC	0.0025 (0.0189)	-0.0077 (0.0232)	0.0030 (0.0304)	0.0099 (0.0124)	0.0027 (0.0145)	0.0135 (0.0186)
Company fixed effects	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included
Constant	-0.4506 (0.3564)	0.4262^{+} (0.2515)	1.3984*** (0.3172)	-0.5417* (0.2284)	0.2145 (0.1642)	1.1262*** (0.2238)

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

In contrast, for posts that did not receive Reactions, the treatment effects on Likes and Comments are negative across all three quantiles. All of these results are directionally consistent with our prior findings. We note that treatment effects for posts that did not receive Reactions are not statistically significant, possibly due to the relatively small sample size in the short-term.

Table 3.16. Long-Term Quantile Regression Estimation Results

DV = Likes						
	Posts with Reactions			Posts without Reactions		
	90%	95%	99%	90%	95%	99%
After	**0.3161***** **(0.0460)**	**0.2281***** **(0.0555)**	**0.3115***** **(0.0915)**	**-0.6586***** **(0.0265)**	**-0.7349***** **(0.0294)**	**-0.7192***** **(0.0556)**
Length	0.0012*** (0.0001)	0.0011*** (0.0001)	0.0009*** (0.0001)	0.0009*** (0.0001)	0.0007*** (0.0001)	0.0006*** (0.0001)
Positive	0.0120*** (0.0014)	0.0115*** (0.0018)	0.0132* (0.0053)	0.0112*** (0.0011)	0.0103*** (0.0012)	0.0116*** (0.0024)
Negative	0.0381*** (0.0019)	0.0285*** (0.0021)	0.0176*** (0.0036)	0.0389*** (0.0014)	0.0321*** (0.0017)	0.0187*** (0.0029)
UGC	0.0007*** (0.0000)	0.0007*** (0.0000)	0.0004*** (0.0000)	0.0006*** (0.0000)	0.0007*** (0.0000)	0.0006*** (0.0001)
MGC	-0.0482*** (0.0050)	-0.0440*** (0.0052)	-0.0151 (0.0111)	-0.0460*** (0.0040)	-0.0465*** (0.0041)	-0.0131 (0.0083)
Month fixed effects	Included	Included	Included	Included	Included	Included
Company fixed effects	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included
Constant	0.6263*** (0.0747)	1.2712*** (0.0708)	2.1879*** (0.0998)	0.6855*** (0.0602)	1.3158*** (0.0602)	2.1272*** (0.0734)
DV = Comments						
	Posts with Reactions			Posts without Reactions		
	90%	95%	99%	90%	95%	99%
After	**0.9837***** **(0.0334)**	**0.8046***** **(0.0424)**	**0.4869***** **(0.0674)**	**-0.2106***** **(0.0240)**	**-0.2469***** **(0.0260)**	**-0.2981***** **(0.0408)**
Length	0.0013*** (0.0001)	0.0015*** (0.0001)	0.0015*** (0.0001)	0.0013*** (0.0001)	0.0013*** (0.0001)	0.0014*** (0.0001)
Positive	-0.0318*** (0.0017)	-0.0238*** (0.0018)	-0.0140*** (0.0025)	-0.0327*** (0.0014)	-0.0251*** (0.0016)	-0.0168*** (0.0015)
Negative	0.0135*** (0.0019)	0.0123*** (0.0021)	0.0194*** (0.0047)	0.0096*** (0.0014)	0.0100*** (0.0016)	0.0110** (0.0037)
UGC	-0.0001** (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
MGC	0.0029 (0.0047)	0.0008 (0.0048)	-0.0082 (0.0073)	0.0037 (0.0039)	-0.0012 (0.0041)	-0.0023 (0.0058)
Month fixed effects	Included	Included	Included	Included	Included	Included
Company fixed effects	Included	Included	Included	Included	Included	Included
Type fixed effects	Included	Included	Included	Included	Included	Included
Constant	-0.3842*** (0.0704)	0.2437*** (0.0674)	1.1317*** (0.0963)	-0.5508*** (0.0613)	0.1535* (0.0627)	1.0124*** (0.0801)

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors included in parentheses.

The long-term regression results in Table 3.16 again present a consistent picture. For posts that received Reactions, we observe a significantly positive effect across all three quantiles. For posts that did not receive Reactions, instead, we observe a significantly negative effect across all three quantiles.

Overall, this set of robustness check results further demonstrated the validity of our previous findings. In particular, the effects of Reactions on Likes and Comments not only exists “on average”, but can also be detected across several quantiles.

3.6 Discussion

User engagement with various content generated on social media takes place via engagement features, such as Likes and Comments. In this essay, we study a new engagement feature on Facebook, known as the Reactions feature, and its impact on existing engagement features. The Reactions feature allow users to express more granular emotions and opinions that cannot be appropriately expressed by Likes, while also keeping the cost of engagement lower than writing Comments.

Although Reactions, by design, seem to fulfill a unique type of engagement need, our results show that Facebook users’ use of the Reactions feature was not independent of other existing engagement features, such as Likes and Comments. By comparing user posts created immediately before the feature change with user posts created immediately after the feature change, we found that, overall, the introduction of Reactions feature increased the number of Likes and decreased the number of Comments that an average user post received.

We further discovered that the overall effects of Reactions on Likes and Comments were in fact shaped by two opposite forces. User posts that received Reactions also ended up receiving more Likes and more Comments than what they would have before the feature change, which provided evidence for a reinforcement relationship between Reactions and existing engagement features. In contrast, posts that were created after the change and that did not receive any Reactions ended up receiving fewer Likes and fewer Comments than what they would have received before the feature change. In other words, the new Reactions feature seemed to lead posts with engagement to get even more engagement. Our analyses also showed that these effects persisted in the long term, at least six months after the introduction of the new feature.

We conducted multiple robustness checks and additional analyses to further establish the validity of our findings. First, our major results can be consistently replicated under several different empirical strategies, including generalized linear regressions, matching methods, as well as quantile regressions. Second, by considering the potential simultaneous use of Reactions and Comments by the same users, we showed that the increase in number of Comments for posts with Reactions was not driven by the same users clicking Reactions buttons and writing Comments at the same time. Third, through an exploratory content analysis of the Comments associated with user posts, we provide initial evidence that, even if there is a co-existence of reinforcement and substitution effects of the Reactions feature, the substitution effect is likely to be quite weak.

One possible mechanism that may explain our findings is a process of "attention redistribution". Because users have limited time and attention to spare on Facebook business pages, they may choose to engage with content that has already received engagement, by using existing engagement as a signal of the underlying quality of the content. As a result, user posts with

Reactions may be perceived as more engaging than user posts without Reactions and subsequently attract even more engagement of various forms. Meanwhile, because users spend more attention on posts with Reactions, less attention is spared on posts without Reactions, leading these posts to receive fewer engagement overall. In summary, now that the effect of the new engagement feature introduction has been robustly established (as was the goal of our study), identifying and pinpointing the exact causal mechanism represents a promising and interesting next step in this stream of research.

This essay contributes novel insights to the social media and online engagement literature by providing empirical evidence on the interplay of multiple engagement features. We demonstrate that usage of different engagement features are not independent from each other. The introduction of a new engagement feature not only provides the users with a new way of expressing themselves and interacting with others, but also changes the usage of existing engagement features. Therefore, future research on social media engagement should take into account the interdependency among engagement features. Furthermore, we show that the new engagement feature does not affect all content equally. Instead, engagement intensity increases for some content but decreases for other content, depending on whether the content has received the new type of engagement (the Reactions in our context). It is important to understand and document not only the short-term impact but also the long-term impact of social media design features, and we show that the effects of engagement feature change can persist over the long term. Our work represents one of the first steps towards understanding the dynamics of engagement behaviors associated with multiple engagement features on social media platforms.

While our work is conducted in the context of Facebook business pages, we believe the key findings are also informative outside Facebook. In particular, most modern social media platforms, including Facebook, offer multiple engagement features to their users. Similar to what we have found on Facebook business pages, the usage of different features is likely to be inter-related on other platforms too, as users utilize the amount of engagement as a heuristic to guide their engagement behaviors. The specific relationships among different features will depend on characteristics of both the platforms as well as the features, but researchers studying online engagement behaviors should take into account the dynamics and inter-dependency of engagement features.

Our findings have practical implications for social media platforms as well as the companies that use social media to engage customers and build their brand communities. For social media platforms, it is useful to be cognizant about the behavioral consequences of a new design feature and to understand how users actually use the new feature. This can help the platforms with designing meaningful new features and properly measuring their efficacy. For the companies, our results indicate that user-generated content with existing engagement is likely to attract even more engagement. Such engagement reinforcement may pose a challenge to the companies, especially when the focal user-generated content is unfavorable toward the companies (e.g., customer complaints or negative reviews). Therefore, companies may want to prioritize in addressing and responding to user content that has received engagement, because such content, both positive and negative, is likely to attract the attention and engagement of more users, and can have far-reaching impact on the users involved.

Chapter 4. Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining

4.1 Introduction

The application of data mining[22] methods creates appealing opportunities for research across multiple disciplines, such as information systems (IS), marketing, economics, and finance. The increasing availability of big data and unstructured data further contributes to the popularity of data mining methods (Agarwal and Dhar 2014; Chen et al. 2012; Varian 2014). Based on observed data, predictive data mining models can be used to automatically generate or estimate variables that researchers are interested in, making it an efficient and sophisticated approach to processing large amounts of structured and unstructured data. Recent examples include the use of text mining techniques to determine the sentiment of text (e.g., Pang et al. 2002; Das and Chen 2007), and the use of image classifiers to predict an individual's gender or race from a profile picture (e.g., Chan and Wang 2014; Rhue 2015), or to detect the presence (absence) of various objects in AirBNB property listings (Zhang et al. 2016).

Many IS studies have recently sought to combine data mining approaches with traditional statistical analyses or econometric modeling in a two-stage process. In the first stage, pre-trained data mining models are deployed to generate new variables that are not readily available from existing data. In the second stage, these generated variables are added into regression models, usually as independent regressors. Several papers adopting this two-stage process have uncovered

[22] We use the general term "data mining" throughout the essay, although the same methodologies are also referred to as "machine learning", "statistical learning", or "predictive analytics" in various contexts.

interesting insights and have been published in top IS journals (e.g., Gu et al. 2007, 2014; Aggarwal et al. 2012; Lu et al. 2013; Moreno and Terwiesch 2014; Wang et al. 2013; Archak et al. 2011). For instance, Aggarwal et al. (2012) adopted a text classification model to label sentiments of online blog posts as positive, negative, and neutral. They then estimated a regression to demonstrate the effect of message sentiment on venture financing outcomes.

However, an important issue with this two-stage process is that variables generated in the first stage almost certainly contain some amount of *predictive error*, because predictive data mining models are imperfect. Such error then carries over to the second stage econometric models, and manifests as *measurement error,* if the variable is continuous, or *misclassification*, if the variable is discrete. For example, suppose we have built a text classification model on a training dataset, which predicts the sentiment of Facebook posts as either positive or negative, and the model has achieved a recall, or sensitivity, of 0.8 for the "positive" class on a holdout, testing dataset. This means that 20% of posts that are actually positive are incorrectly classified as negative. These errors, if ignored, can introduce systematic biases into the second stage estimations and may, therefore, threaten the validity of the subsequent statistical inferences.

The issue of measurement error and misclassification is not new and has received a great deal of attention from econometricians and statisticians (Greene 2003). However, it warrants special attention in the new context of big data and increasing interest in combining data mining with econometric modeling for the following reasons. First and foremost, measurement error is unobservable in many situations; however, here the errors, which originate from imperfect predictions by first stage data mining models, can be observed and quantified using standard methods of model evaluation, stemming from confusion matrices or continuous measures of error. This provides a clear opportunity to diagnose the error and correct for the resulting bias. Second, many if not most studies in IS that have used the two-stage approach of combining econometric modeling with data mining have failed to acknowledge the potential estimation biases introduced by measurement error or misclassification. We believe that this may derive, at least in part, from a lack of understanding or awareness of the issue. Third, the variables obtained from the first-stage prediction typically enter the second-stage estimation as independent regressors. Unlike error in dependent variables, which typically leads to inflated variance of estimates and decreased model fit, error in independent variables generally introduces systematic biases into coefficient estimates (Greene 2003), and thus causes serious concerns.[23] Yet, most IS researchers seem to be unaware of either the potential biases from predictive errors or proper methods to mitigate the biases.

In this essay, we hope to bridge this gap by addressing three key issues: (1) To what extent will measurement error or misclassification from data mining models bias estimations in econometric analyses that incorporate the output of those models? (2) How can we diagnose the structure of the measurement error or misclassification, and the resulting biases, in a particular research setting and dataset? (3) How can we mitigate these biases?

Based on both theoretical reasoning and simulated data, we first demonstrate that

[23] In this essay, we do not consider the issue of error in dependent variables, because it is rare for studies to employ predictive models to generate outcome variables for second stage estimations. Indeed, during our review of the literature for this essay, we did not come across any study in the IS literature that has taken this approach.

measurement error and misclassification can indeed introduce considerable biases into several commonly used econometric models, such as linear regressions, generalized linear regressions (e.g., Logit, Probit, and Poisson models), and panel data regressions. Notably, our simulations are conducted based on commonly observed levels of predictive performance in data mining models, in terms of error variance for numeric predictions or precision and recall for classifications. Hence, the errors we simulate and the biases we observe are likely to manifest in an actual study.

Having established the undesirable impact of error on econometric analyses, we then review several possible error-correction methods. We focus on two simulation-based methods that lend themselves well to mitigating the bias introduced by predictive measurement error and misclassification. The Simulation-Extrapolation (SIMEX thereafter) method applies to continuous variables with additive measurement error (Cook and Stefanski 1994). The Misclassification-SIMEX (MC-SIMEX thereafter) method applies to discrete variables with misclassification (Küchenhoff et al. 2006). We focus on SIMEX and MC-SIMEX rather than other approaches such as instrumental variable approach or method-of-moments for two main reasons. First, SIMEX and MC-SIMEX can easily be applied to a variety of model specifications whereas most other methods require model-specific assumptions. Second, SIMEX and MC-SIMEX can be configured based solely on the observable performance indicators of first-stage data mining models, whereas other methods typically require explicit modeling of errors in the second-stage estimations. We validate the effectiveness of SIMEX and MC-SIMEX using simulated data, and we then apply both methods to three real world datasets. Our results demonstrate the effectiveness of these methods in mitigating estimation bias from measurement error and misclassification. Our results also reveal the limitation of these or any methods in addressing predictive measurement error issues, when first stage data mining performance is problematically low. Finally, we provide a guiding procedure that researchers can follow to diagnose estimation biases and assess the efficacy of specific error correction methods in consideration of their research settings, with specific data samples and data mining models.

This essay contributes to the IS literature in three ways. First, we describe and raise awareness of the issue of measurement error and misclassification in the context of an increasingly prevalent methodological practice in IS research, i.e., the integration of data mining and econometric analyses. We show that, while predictive error can bias econometric estimations, the ability to quantify such error brings the opportunity to correct for estimation biases. Second, we review several existing remedial approaches that can address the identified issue, and demonstrate the effectiveness of two methods in particular, using both simulations and real-world empirical applications. Third, we propose a diagnostic procedure via which researchers can assess the characteristics of the measurement error and estimation bias in a particular scenario, with a given sample of data, and thereby choose the best approach to address the problem in that setting.

Measurement error and misclassification may arise in a variety of research settings and are very difficult to avoid completely. Therefore, we believe that awareness of the problem and the severity of its consequences can help researchers understand the potential risks of combining data mining with econometric analyses, and thus to improve the robustness of their conclusions. At the same time, we stress that the points raised in this essay do not necessarily invalidate the results of any past work, because the predictive error in the first stage data mining can have variable effects on the subsequent econometric estimation. The predictive error may cause attenuation of coefficients in some cases, amplification in others, and in some cases it may have little effect at all. Thus, our aim with this essay is to highlight the unique opportunity of error correction in this setting and to provide IS scholars with guidance on the implementation of this integrated methodology in as robust a manner as possible, going forward.

4.2 The Common Practice of Combining Data Mining and Econometric Analyses

Studies that have adopted the two-stage methodology of combining data mining techniques with econometric estimations are becoming prevalent in the IS discipline. A cursory search of recently published issues of top IS journals and conference proceedings revealed at least 13 studies that have used this approach; we identified 6 recent studies in *Information Systems Research* (Gu et al. 2007, 2014; Aggarwal et al. 2012; Wang et al. 2013; Moreno and Terwiesch 2014; Singh et al. 2014), 2 in *Management Science* (Archak et al. 2011; Lu et al. 2013), 2 appearing in other journals (Ghose and Ipeirotis 2011; Ghose et al. 2012), and 3 in the *Proceedings of the International Conference on Information Systems* (Chan and Wang 2014; Rhue 2015; Zhang et al. 2016).[24] The two-stage methodology has also been adopted in several fields outside the IS community, such as Marketing

[24] We searched for papers that used predictive data mining methods (e.g., classification) and excluded studies that only employed dictionary-based natural language processing techniques (e.g., Johnson et al. 2015; Tetlock et al. 2008) and studies that used exploratory data mining methods (e.g., Wu 2013; Bao and Datta 2014). In this essay, we do not discuss exploratory data mining models, such as topic modeling using Latent Dirichlet Allocation (LDA), because they generally do not have prediction-oriented evaluation metrics that can be used to make error corrections.

(e.g., Tirunillai and Tellis 2012), Human-Computer Interaction (e.g., Liu et al. 2012; Zhu et al. 2011, 2012), Economics (e.g., Jelveh et al. 2014), and Finance (see Fisher et al. 2016 for a review). In this section, we report and discuss several patterns we have observed in these publications.

The most common application of data mining models in these studies was text classification that was used primarily for coding online user-generated content, such as consumer reviews. Another, less common use was image classification that was used to identify objects or persons from digital photographs (Ghose et al. 2012; Chan and Wang 2014; Rhue 2015; Zhang et al. 2016). Most of the papers followed the common approach to develop the classification models.[25] To build a classification model, researchers first draw a random subsample of observations from the dataset and have them manually classified or labeled by human coders based on predefined rules. This manually classified subsample then becomes the ground truth for training and evaluating the classifier.[26] A classifier is trained using a portion of the labeled data and then its performance is evaluated using the remaining data, by comparing the classifier's predictions with the ground truth. Some studies (e.g., Ghose and Ipeirotis 2011; Ghose et al. 2012) have adopted a more

---

[25] For a comprehensive introduction to data mining or textual classification, readers may refer to Aggarwal (2015) or Provost and Fawcett (2013). Varian (2014) also provides an overview of data mining techniques for econometricians.

[26] In this essay, we focus on predictive errors from data mining models. We do not consider inter-coder disagreement or error introduced via the human-labeling process. We believe that disagreements amongst human coders are fundamentally different from predictive errors. Manual labeling is most often employed when there is no ground truth. Disagreements among coders typically reflect inherent ambiguity or subjectivity in the coding process, whereas predictive errors typically reflect the limited learning capacity of data mining models. For subjective or open-ended labeling tasks, the issue of coder-introduced error might be less concerning because the labels reflect researchers' subjective belief about "ground truth" and may not contain definitive error. The application of data-driven procedures to resolve inter-coder disagreement falls outside the scope of this work. However, for an example that discusses the issue of inter-coder disagreement and the use of SIMEX to mitigate its impact, see Hopkins and King (2010).

advanced evaluation method, known as cross-validation, wherein the labeled set is partitioned into K folds, and classifiers are iteratively trained on different sets of (K-1) folds and evaluated using the remaining fold. The trained classifiers are then deployed on the unlabeled remainder of the dataset to obtain predicted labels. This approach has the benefit of scalability, because hand-coding an overwhelmingly large dataset is often infeasible.

There exist many data mining techniques for building predictive models, including classification and regression trees, k-nearest neighbors, naïve Bayes, neural networks, support vector machines, Bayesian networks, and various linear and non-linear regression techniques. Some of the techniques were developed to predict continuous outcomes (numeric prediction task), some to predict discrete outcomes (classification task), and others can be configured for either purpose. Several metrics are available to assess their predictive performance. For numeric prediction, evaluations are based on prediction errors, i.e., the differences between predicted and actual values. Commonly used metrics include MAE (mean absolute error) and RMSE (root mean squared error) (Aggarwal 2015). For classification models, commonly used metrics include overall accuracy (the percentage of correct predictions across all classes), precision (the percentage of predictions in a given class that are correct), and recall (the percentage of cases that truly belong to a given class that are correctly predicted by the model) (ibid). Figure 4.1 illustrates these performance metrics using a binary classification model as an example. All papers we surveyed used classification models; 6 papers reported the predictive performance of their data mining models, with overall accuracy ranging from 60% to 87%, precision ranging from 70% to 100%, and recall ranging from 74% to 100%.

Another pattern we observed was that, in all papers, the variables generated via data mining were incorporated into second-stage regressions with many other covariates. Typical second-stage econometric models include linear regressions with fixed or random effects, Logit or Probit regressions, systems of equations, and vector autoregression (VAR). We also observed that the econometric models were typically estimated on a much larger sample than the one used to train the data mining model in the first stage. For example, Moreno and Terwiesch (2014) used a labeled sample to train their model that comprised 2% of the total dataset. The trained model was then used to generate the variable of interest for the remaining 98% of the dataset. As we will show in the next section, the measurement error or misclassification that originates from data mining has the potential to introduce systematic biases into subsequent econometric estimations. These biases persist in large samples, and are generally harder to anticipate or predict as specification of econometric model grows more complex.

Figure 4.1. Performance Metrics for a Two-Class Classification Model

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Accuracy = (TP + TN)/(TP + TN + FP + FN)
For positive class:
Precision = TP/(TP + FP),
Recall = TP/(TP + FN)
For negative class:
Precision = TN/(TN + FN),
Recall = TN/(TN + FP)

Note. The left-hand panel is a confusion matrix obtained by evaluating a given predictive model. It summarizes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The right-hand panel lists the performance metrics derived from the confusion matrix, including overall accuracy, precision, and recall rates.

4.3 Estimation Biases due to Measurement Error or Misclassification

In this section, we present both analytical and simulation results regarding the biases in coefficient estimates caused by measurement error or misclassification, for several commonly used econometric models. First, we discuss a simple linear regression with one regressor, containing either measurement error or misclassification. In this scenario, the bias can be mathematically derived. Subsequently, for more complicated model specifications, we demonstrate the resultant biases using simulated data.

4.3.1 Bias in linear regression with one regressor

Consider a simple linear regression with only one regressor: $Y = \beta_0 + \beta_1 X + \varepsilon$. If both the dependent and independent variables are precisely measured, OLS would yield unbiased, consistent, and efficient estimates of β_0 and β_1 (Greene 2003). Now, suppose that instead of X we actually observe $\hat{X}$, which includes error. Regressing Y on $\hat{X}$ would yield biased estimates.

If X is a continuous variable, there are two broad types of measurement error: *classical* error and *non-classical* error. If the measurement error, e, is *random* and *additive* – i.e., $\hat{X} = X + e$ – and *independent* of both X and ε, such error is known as classical measurement error (Carroll et al. 2006). The error results in an *attenuation bias*, that is, the estimated $\widehat{\beta_1}$ satisfies $E(\widehat{\beta_1}|\hat{X}) = \beta_1[\sigma_X^2/(\sigma_X^2 + \sigma_e^2)]$, which implies that the regression coefficient is underestimated (see Greene 2003 for proof). Given X, the magnitude of the bias depends on the variance of the error, and larger error variance leads to greater bias. Measurement error that is not random, not additive, or not independent of X and ε is known as non-classical measurement error, and we will discuss its impact later.

If X is a discrete variable, the misclassification would also result in a systematic bias in the regression coefficient. For simplicity, we can assume $\hat{X}$ is a dummy variable and conditionally independent of Y given X (i.e., nondifferential misclassification),[27] then the estimated $\widehat{\beta_1}$ will satisfy $E(\widehat{\beta_1}|\hat{X}) = \beta_1[Pr(X=1|\hat{X}=1) - Pr(X=1|\hat{X}=0)]$ (Gustafson 2003, see Appendix 4.1 for proof). Using data mining performance measures, this relationship can be written as follows: $E(\widehat{\beta_1}|\hat{X}) = \beta_1[Pr(X=1|\hat{X}=1) + Pr(X=0|\hat{X}=0) - 1] = \beta_1[Precision(class\ 1) + Precision(class\ 0) - 1]$. That is, the magnitude of the bias is determined by the sum of the precision rates for the two classes. Appendix 4.2 provides an example and a graphical illustration of how misclassification can result in estimation bias. In extreme cases when the sum of the two precision scores is smaller than 1, the estimated coefficient may shift in the opposite direction from the true value, resulting in a coefficient of the opposite sign.

Finally, note that the above finite sample results also hold asymptotically. For continuous measurement error, $plim\,\widehat{\beta_1} = \beta_1[\sigma_X^2/(\sigma_X^2 + \sigma_e^2)]$. For binary misclassification, $plim\,\widehat{\beta_1} = \beta_1[Pr(X=1|\hat{X}=1) - Pr(X=1|\hat{X}=0)]$. Therefore, coefficient estimates with errors are inconsistent.

4.3.2 Bias in more complicated models: Theoretical results

Considering the above discussion, one might be tempted to conclude that measurement error and

[27] This assumption is likely to hold if $\hat{X}$ is generated via a data mining model, because $\hat{X}$ is only determined by its true value, X, and the data mining model, which is usually a separate process from the data-generating process reflected by the regression equation.

misclassification will typically only produce an *attenuation bias* in coefficient estimates and, thus, will only lead to conservative results and Type II error. However, it is important to note that an *amplification bias* may also manifest. This can happen when either the error structure or the econometric specification in the second-stage regression model grows more complicated.

First, in the case of linear regression with one regressor, amplification bias can manifest under non-classical measurement error whose error structure deviates from the random, additive, and independent error that we described in Section 4.3.1. For example, consider a continuous variable with an additive measurement error, where the error structure includes both a random component and a systematic component in the form of $\hat{X} = a + bX + e$, $E(e) = 0$, where a represents the additive systematic error, b represents the multiplicative systematic error, and e represents the random error. The resulting coefficient on $\hat{X}$ then satisfies $E(\widehat{\beta_1}|\hat{X}) = \frac{\beta_1 b \sigma_X^2 + \rho_{e\varepsilon}\sigma_e\sigma_\varepsilon}{b^2\sigma_X^2 + \sigma_e^2}$ (Carroll et al. 2006). Attenuation bias happens with a classical measurement error, as we illustrated in the previous section, only because we assumed (1) e is uncorrelated with ε, the regression error term (i.e., $\rho_{e\varepsilon} = 0$), and (2) there is no systematic error between X and $\hat{X}$, i.e., $b = 1$. In our scenario of interest, the form of the measurement error is determined by the data mining model. If the data mining model systematically underestimates the true value, i.e., $b < 1$, then the second assumption is no longer true, and the bias in β_1 may manifest as an amplification. An amplification bias can occur in misclassification as well. Gustafson (2003) notes that, if a categorical variable with more than two levels bears misclassification, no simple conclusion can be drawn about the direction of the bias in the estimated coefficient.

Second, when the second-stage regression specification becomes more complicated, the biases can be similarly difficult to anticipate. In multivariate regressions, even when the other variables (i.e., those not generated from data mining) are measured without error, the presence of a data mined variable with predictive error can cause the coefficient estimates of all variables to be biased in unknown directions (Greene 2003; Gustafson 2003; Buonaccorsi 2005). In nonlinear regressions, the directions of biases associated with both the variable with error and the other precisely measured variables are also uncertain (Carroll et al. 2006).

4.3.3 Bias in more complicated models: Simulation results

Because analytical, closed form solutions are generally difficult to obtain for complicated regression models with measurement error or misclassification, we provide an illustrative numerical analysis based on simulation. The simulation was conducted as follows. First, we generated three variables having different underlying distributions: $X_1 \sim N(0, 1^2)$, $X_2 \sim Bernoulli(p)$, and $X_3 \sim Uniform(-10,10)$. We modified X_1 and X_2 to introduce measurement error or misclassification. Second, we generated another normally distributed variable as the error term as $\varepsilon \sim N(0,0.5^2)$. Third, we generated a dependent variable as a function of the independent variables and the error term: $Y = 1 + 2 \times X_1 + 3 \times X_2 + 0.5 \times X_3 + \varepsilon$. The coefficients were fixed in order to quantify the magnitudes of estimation biases. In addition to linear regression, we also simulated Logit, Probit, and Poisson regressions, as well as a linear regression with fixed effects. For the three generalized linear models, we generated dependent variables based on the corresponding distributional assumptions.[28] For the linear panel data model with fixed effects, the regression estimated was $Y_{ij} = \alpha_i + 2 \times X_{1ij} + 3 \times X_{2ij} + 0.5 \times X_{3ij} + \varepsilon_{ij}$, where $\alpha_i = i, i \in \{1,2, \ldots ,25\}$ and $j \in \{1,2, \ldots ,200\}$. The α_i represented the panel-specific fixed effects. We also simulated a linear random-effects regression, where α_i were randomly drawn from a standard normal distribution. The results were qualitatively the same as the fixed-effect model, so we only reported the fixed-effect regression. The results that we report below are based on 5,000 observations. We repeated the analysis with 10,000 observations and got similar results with no qualitative differences. Below we present three simulation results, respectively showing estimation biases caused by classical and non-classical measurement error in X_1, and misclassification in X_2.

We first simulated the impact of classical additive measurement error in X_1 with $\widehat{X_1} = X_1 + e$, where $e \sim N(0, \sigma_e^2)$ and is independent of X_1, ε, and the other covariates in the simulated regressions. Because X_1 follows a standard normal distribution, we considered three values of σ_e – 0.1, 0.3, and 0.5 – to capture different degrees of measurement error. For all simulations in this part, $X_2 \sim Bernoulli(0.3)$ and contained no misclassification, enabling us to isolate the impact of measurement error in X_1. Table 4.1 summarizes our simulation results. For each regression, the first column shows coefficient estimates without measurement error (denoted as b) and the second column shows coefficient estimates with measurement error in X_1 (denoted as b'). The third column shows the relative magnitude of estimation bias, calculated as % = $(b' - b)/b$. For the linear fixed-effects regression, we omit the estimates of the fixed effects, due to space

[28] Let $Xb = 1 + 2X_1 + 3X_2 + X_3$. Y_{Logit} is drawn from a Bernoulli distribution with $p = \frac{1}{1+e^{-Xb}}$. Y_{Probit} is drawn from a Bernoulli distribution with $p = \phi(Xb)$. $Y_{poisson}$ is drawn from a Poisson distribution with $\lambda = e^{Xb}$.

consideration. Because the data was simulated, all estimates were statistically significant. We therefore do not report standard errors or levels of statistical significance.

Table 4.1. Regression Results for X_1 with Classical Measurement Error

	OLS			Logit			Probit			Poisson			Fixed-Effect		
	b	b'	%	b	b'	%	b	b'	%	b	b'	%	b	b'	%
$\sigma_e = 0.1$															
C	1.004	1.006	0.3%	1.002	0.993	-0.9%	1.042	1.012	-2.8%	0.999	1.232	23%			
X_1	1.995	1.970	-1.3%	1.996	1.954	-2.1%	1.979	1.903	-3.9%	2.000	1.911	-4.4%	1.994	1.977	-0.9%
X_2	2.988	2.990	0.1%	2.890	2.868	-0.8%	2.897	2.822	-2.6%	3.000	2.956	-1.5%	2.986	2.987	0.03%
X_3	0.499	0.500	0.2%	0.493	0.489	-0.8%	0.484	0.471	-2.7%	0.500	0.486	-2.8%	0.499	0.499	0%
$\sigma_e = 0.3$															
C	1.004	1.005	0.2%	1.002	0.962	-3.9%	1.042	0.911	-12.5%	0.999	1.482	48%			
X_1	1.995	1.833	-8.1%	1.996	1.748	-12.4%	1.979	1.567	-20.8%	2.000	1.760	-12.0%	1.994	1.824	-8.5%
X_2	2.988	2.980	-0.3%	2.890	2.757	-4.6%	2.897	2.499	-13.7%	3.000	2.866	-4.5%	2.986	2.990	0.13%
X_3	0.499	0.500	0.2%	0.493	0.472	-4.3%	0.484	0.421	-13.2%	0.500	0.474	-5.1%	0.499	0.496	-0.6%
$\sigma_e = 0.5$															
C	1.004	0.996	-0.8%	1.002	0.899	-10.2%	1.042	0.765	-26.6%	0.999	2.445	144%			
X_1	1.995	1.595	-20.0%	1.996	1.453	-27.2%	1.979	1.155	-41.7%	2.000	1.337	-33.1%	1.994	1.589	-20.3%
X_2	2.988	3.011	0.8%	2.890	2.678	-7.3%	2.897	2.205	-23.9%	3.000	2.750	-8.3%	2.986	2.983	-0.1%
X_3	0.499	0.500	0.2%	0.493	0.445	-9.7%	0.484	0.357	-26.2%	0.500	0.418	-16.4%	0.499	0.499	0%

Note. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_1. % stands for relative magnitude of estimation bias.

Several patterns emerged that are worth noting. While the coefficient on X_1 was consistently downward biased, coefficients of other variables were biased in different directions. As the magnitude of measurement error increased from 0.1 to 0.5, bias in the coefficient of X_1 also increased from -1.3% to -20% in the case OLS. Compared to OLS, biases in generalized linear models were greater. For example, with measurement error of $\sigma_e = 0.5$, the coefficient of X_1 in OLS was underestimated by 20%, compared to 27.2% in Logit, 41.7% in Probit, and 33.1% in Poisson regression. Bias in the linear fixed-effect model was comparable to bias in OLS, and estimates of the fixed effects were unbiased.

Next, we simulated three types of non-classical measurement error in X_1: (1) $\widehat{X_1} = X_1 + e,\ e \sim N(0, \sigma_e^2)$ and is independent of ε and other covariates, but is correlated with the true value X_1 with $\rho_{X_1 e} = 0.5$; (2) $\widehat{X_1} = X_1 + e,\ e \sim N(0, \sigma_e^2)$ and is independent of X_1 and ε, but is correlated with X_3 with $\rho_{X_3 e} = 0.5$; (3) $\widehat{X_1} = 1 + 0.5X_1 + e,\ e \sim N(0, \sigma_e^2)$ and is independent of X_1, ε, and other covariates. The first two scenarios represent random measurement error that is correlated with either the true value or another covariate in the second-stage regression, and the third scenario represents systematic independent measurement error. All three scenarios of error may occur in data mining model predictions. For simplicity, we only report simulation results for linear regressions in Table 4.2. We obtained similar results for other regressions.

Table 4.2. Regression Results for X_1 with Non-Classical Measurement Error

		Scenario (1)			Scenario (2)			Scenario (3)		
		b	b'	%	b	b'	%	b	b'	%
$\sigma_e = 0.1$	C	1.004	1.003	-0.1%	1.004	1.005	0.1%	1.004	-2.817	-380%
	X_1	1.995	1.884	-5.6%	1.995	1.979	-0.8%	1.995	3.827	91.8%
	X_2	2.988	2.983	-0.2%	2.988	2.983	-0.2%	2.988	2.977	-0.4%
	X_3	0.499	0.498	-0.2%	0.499	0.481	-3.6%	0.499	0.498	-0.2%
$\sigma_e = 0.3$	C	1.004	1.000	-0.4%	1.004	1.006	0.2%	1.004	-1.939	-293%
	X_1	1.995	1.641	-17.7%	1.995	1.863	-6.6%	1.995	2.903	45.5%
	X_2	2.988	2.976	-0.4%	2.988	2.975	-0.4%	2.988	3.006	0.6%
	X_3	0.499	0.498	-0.2%	0.499	0.450	-9.8%	0.499	0.500	0.2%
$\sigma_e = 0.5$	C	1.004	0.997	-0.7%	1.004	1.005	0.1%	1.004	-0.971	-197%
	X_1	1.995	1.412	-29.2%	1.995	1.667	-16.4%	1.995	1.941	-2.7%
	X_2	2.988	2.973	-0.5%	2.988	2.971	-0.6%	2.988	2.995	0.2%
	X_3	0.499	0.497	-0.4%	0.499	0.425	-14.8%	0.499	0.498	-0.2%

Note. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_1. % stands for relative magnitude of estimation bias.

Several patterns emerged that are worth noting. Under scenario (1), where measurement error was correlated with the true value of X_1, we observed greater downward bias in the coefficient on X_1 than the bias from classical measurement error. Under scenario (2), where the error was correlated with X_3, we observed biases in the coefficients of both X_1 and X_3. Under scenario (3), where the measurement error was systematic, we observed overestimation of the coefficient of X_1 for $\sigma_e = 0.1$ and $\sigma_e = 0.3$, but attenuation for $\sigma_e = 0.5$. In other words, as error variance became greater, the bias shifted from amplification to attenuation. As noted previously, this result demonstrates numerically that measurement error introduced during first-stage data mining tasks do not necessarily result in attenuation and conservative estimates; in some cases, it may result in *amplified* coefficient estimates.

Finally, we simulated misclassification by modifying the value of X_2. We use a misclassification matrix to represent the magnitude of misclassification in X_2.[29] For a binary variable, the misclassification matrix can be denoted as $(M_{00}, M_{10}, M_{01}, M_{11})$, where $M_{ab} = Pr(\widehat{X_2} = b | X_2 = a)$. It can also be written, equivalently as $(M_{00}, 1 - M_{11}, 1 - M_{00}, M_{11})$, where M_{00} is the recall rate for class 0 (true negative rate) and M_{11} is the recall rate for class 1 (true positive rate). We generate $\widehat{X_2}$ by adjusting the value of X_2, changing it from 0 to 1 with a probability of M_{01} and from 1 to 0 with a probability of M_{10}. Using this method, $\widehat{X_2}$ simulates predicted values from a binary classifier, with recall rate M_{00} for class 0 and recall rate M_{11} for class 1.

To examine the impact of different levels of misclassification, we simulated three scenarios: (1) $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.8$, and $M_{11} = 0.8$; (2) $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.6$, and $M_{11} = 0.5$; and (3) $X_2 \sim Bernoulli(0.5)$, $M_{00} = 0.6$, and $M_{11} = 0.5$. Scenarios (1) and (2)

[29] The misclassification matrix, while different from a confusion matrix, is readily constructed from the confusion matrix by calculating the recall rates for each class.

had a skewed Bernoulli distribution for the true value of X_2, with $Pr(X_2 = 1) = 0.3$; and scenario (3) had a balanced distribution. Scenario (2) and (3) also had greater misclassification than scenario (1). For all simulations in this part, $X_1 \sim N(0, 1^2)$ and contained no measurement error. Table 4.3 summarizes the results.

Table 4.3. Regression Results for X_2 with Misclassification

	OLS			Logit			Probit			Poisson			Fixed-Effect		
	b	b'	%	b	b'	%	b	b'	%	b	b'	%	b	b'	%
Scenario (1): $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.8$, and $M_{11} = 0.8$															
C	1.004	1.292	28.7%	1.002	1.163	16.1%	1.042	0.943	-9.5%	0.999	1.699	70.1%			
X_1	1.995	1.995	-0.0%	1.996	1.689	-15.4%	1.979	1.342	-32.2%	2.000	1.878	-6.1%	1.994	2.009	0.75%
X_2	2.988	1.596	-46.6%	2.890	1.106	-61.7%	2.897	0.979	-66.2%	3.000	1.722	-42.6%	2.986	1.583	-47.0%
X_3	0.499	0.494	-1.0%	0.493	0.413	-16.3%	0.484	0.328	-32.3%	0.500	0.533	6.5%	0.499	0.492	-1.4%
Scenario (2): $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.6$, and $M_{11} = 0.5$															
C	1.004	1.768	76.1%	1.002	1.370	36.7%	1.042	1.131	8.6%	0.999	2.671	168%			
X_1	1.995	2.004	0.5%	1.996	1.616	-19.0%	1.979	1.221	-38.3%	2.000	1.831	-8.5%	1.994	2.001	0.35%
X_2	2.988	0.282	-90.6%	2.890	0.304	-89.5%	2.897	0.148	-94.9%	3.000	0.205	-93.2%	2.986	0.271	-90.9%
X_3	0.499	0.493	-1.1%	0.493	0.393	-20.4%	0.484	0.296	-38.9%	0.500	0.536	7.1%	0.499	0.492	-1.4%
Scenario (3): $X_2 \sim Bernoulli(0.5)$, $M_{00} = 0.6$, and $M_{11} = 0.5$															
C	1.002	2.352	135%	0.980	1.744	78.0%	0.917	1.260	37.4%	1.000	3.662	266%			
X_1	1.995	1.987	-0.4%	1.961	1.477	-24.7%	1.927	1.091	-43.4%	2.000	1.823	-8.8%	1.994	1.988	-0.3%
X_2	2.997	0.332	-88.9%	2.923	0.266	-90.9%	2.910	0.243	-91.7%	3.000	0.032	-98.9%	2.999	0.272	-90.9%
X_3	0.499	0.503	0.8%	0.502	0.388	-22.8%	0.475	0.276	-41.7%	0.500	0.480	-4.0%	0.499	0.503	0.8%

Note. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_2. % stands for relative magnitude of estimation bias.

Once again, several patterns emerged that are worth noting. First, even if a classifier achieved a reasonable level of performance in terms of precision and recall, the misclassification could still lead to severe bias in the coefficient estimates. For example, scenario (1) represented a binary classifier with an 80% recall rate for both classes, as well as 63% precision for the positive class and 90% precision for the negative class. Based on the published work we have surveyed, this level of performance would be considered good in many application domains. However, our simulation showed that the coefficient on X_2 was underestimated by 46.6% in the OLS regression. Second, we observed similar biases in scenarios (2) and (3) although the magnitude of the biases was greater than scenario (1). Due to the greater misclassification in scenarios (2) and (3), the coefficient on X_2 was reduced nearly to zero although it remained statistically significant. Third, in the linear fixed-effect model, the estimates of the fixed effects were also biased to various degrees, ranging from 4% to 33% overestimation. Overall, our simulation results demonstrate the biases associated with misclassification, and the risk of making inferences from the resultant estimates.

4.4 Bias Correction

Section 4.3 provides ample evidence that measurement error and misclassification, which can be introduced with the application of data mining techniques, may severely bias the estimates of econometric models. This poses serious challenges to the increasingly prevalent practice of combining data mining with econometric analysis. However, the good news is that, although data mining models produce predictions with error, the standard practice of model performance evaluation affords a readily accessible quantification of the error. Quantifying error allows one to employ corrective methods that can mitigate subsequent estimation biases. In this section, we first review several existing error-correction methods. Then, we focus on two simulation-based methods (SIMEX and MC-SIMEX), which were initially developed in the field of biostatistics and can be used to mitigate bias in second-stage econometric estimations. We describe the general process which researchers can follow to quantify and correct errors in their datasets. We then use simulations to show the effectiveness of SIMEX and MC-SIMEX methods.

4.4.1 Review of Bias Correction Methods

There have been at least five popular bias correction methods discussed in the research literature, including (1) instrumental variables, (2) method-of-moments, (3) likelihood-based methods, (4) regression calibration, and (5) simulation-extrapolation (SIMEX).

The *instrumental variable approach* can be used to address all kinds of endogeneity issues in regression including measurement error. In a linear regression $Y = X\beta + \boldsymbol{Z\gamma} + \varepsilon$, where X contains additive measurement error, i.e., $\hat{X} = X + e$, the regression model can be rewritten as $Y = \hat{X}\beta + \boldsymbol{Z\gamma} + (\varepsilon - e\beta)$. Thus, the variable with error $\hat{X}$ is correlated with the error term, causing endogeneity. If the researcher can find an appropriate instrument, W, that is correlated with $\hat{X}$ but not with the error term, then a two-stage least squares (2SLS) estimator can be used to obtain the unbiased estimate of the coefficient of X.

Alternatively, if the researcher has accurate knowledge about the moments of measurement error and other variables in the econometric model, the unbiased coefficients may be recovered under some specifications, either analytically or numerically. This approach is known as the *method-of-moments approach*, or *functional approach* (Carroll et al. 2006). In linear regressions with only one regressor, this approach is very straightforward. If the values of σ_X^2 and σ_e^2 are known or can be estimated, one can easily calculate the corrected coefficient as $\widehat{\beta_1}[(\sigma_X^2 + \sigma_e^2)/\sigma_X^2]$. In multivariate linear models or nonlinear models, one also needs knowledge of the covariance between the measurement error and other covariates.

Another option is the *likelihood-based method*, which involves explicit modeling of the error, that is, modeling the probability of observing the values of the dependent variable, given the values of the independent variables. Typically, in order to model this likelihood, researchers need to make distributional assumptions, such as the conditional distribution of a variable with error given its true values, and the distribution of the true values (Carroll et al. 2006). If such information is available, then the likelihood-based method can help recover unbiased estimates via maximum likelihood estimation.

There are also data-driven approaches such as *regression calibration*, which is a general-purpose bias correction method (Gleser 1990). Imagine that, for a subset of data, researchers can observe both the variable measured with error ($\hat{X}$) and its true value (X). Using this subset, it is then possible to fit a regression model of X on $\hat{X}$ and the other observed covariates ($\boldsymbol{Z}$), denoted as $f(\hat{X}, \boldsymbol{Z})$. For remaining data where X is not observable, it can be estimated via the model $f(\hat{X}, \boldsymbol{Z})$. Then, using the estimated values of X and other precisely measured covariates, the researcher can carry out the desired econometric analyses under an assumption that no measurement error remains. This method essentially views measurement error as a missing data problem. The true values for the variable with error are considered missing, and are imputed from a predictive model built on the subsample of data where true values are observed.

Finally, another general-purpose, data-driven approach to bias correction is *simulation-extrapolation* or SIMEX (Cook and Stefanski 1994). As a simulation-based method, the SIMEX method has several advantages over the other methods, in dealing with measurement error and misclassification caused by data mining models. Compared to the first three methods outlined above, SIMEX requires relatively little information and fewer assumptions. For example, the instrumental variable approach requires the identification of an appropriate instrument. The method-of-moments approach requires knowledge of the moments of measurement error as well as the covariance between the error and other covariates. The likelihood-based method requires researchers to make distributional assumptions. In contrast, SIMEX requires only information on the variance of measurement error or a misclassification matrix, which is readily available from the performance evaluation measures of data mining models. Both the error variance and misclassification matrix can be calculated by comparing model predictions with true values using the test dataset. Because the test set is typically a random subsample of the labeled data, the calculated error variance or misclassification matrix can be generalized to the broader, unlabeled data.

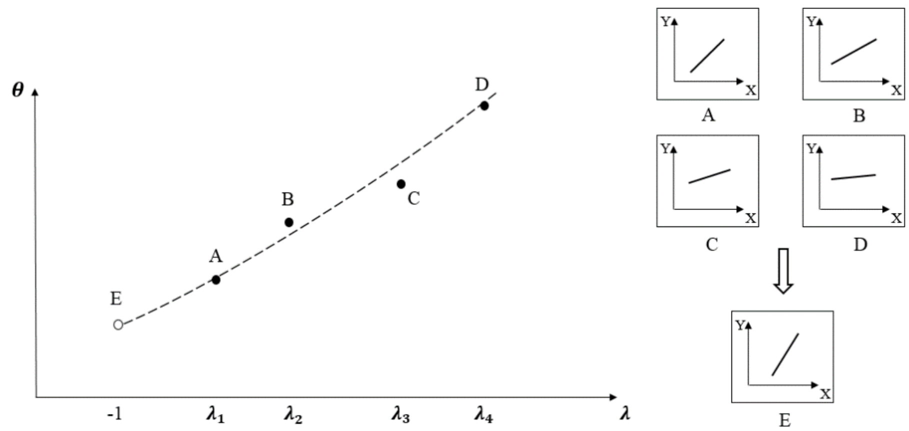
SIMEX demonstrated better performance than regression calibration under a number of scenarios, in particular under nonlinear econometric specifications. We experimented with both regression calibration and SIMEX for Logit, Probit, and Poisson regressions, where one independent variable was normally distributed ($\sigma_e = 0.3$) and contained classical measurement error. The error biased the coefficient estimate from 2 down to 1.748, 1.567, and 1.760 respectively. SIMEX was able to correct the coefficient back to 1.999, 1.878, and 1.977, whereas regression calibration only corrected the coefficient back to 1.973, 1.815, and 1.927. Moreover, SIMEX requires less time and effort to execute because the procedure has been implemented in software packages that are commonly available for statistical analyses. For example, SIMEX is available in

R, an open source statistical programming language, via the *simex* package, and also available in STATA, via the *simex* function (Hardin et al. 2003). Due to the above reasons, we focus on SIMEX as the primary correction method in this essay. Meanwhile, we encourage researchers to consider and evaluate multiple error correction procedures including SIMEX to identify the best fit for their research setting, data, and data mining models, via the diagnostic procedure we outline below in Table 4.4 (Section 4.4.3).

4.4.2 Introduction to SIMEX and MC-SIMEX

The Simulation-Extrapolation (SIMEX) method was proposed by Cook and Stefanski (1994) to address additive measurement error in a *continuous* variable (i.e., $\hat{X} = X + e$) in models where the error variance σ_e^2 is known or can be accurately estimated. The SIMEX method consists of two steps: a simulation step and an extrapolation step. In the simulation step, a fixed set of non-negative values $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is selected (e.g., $\{1,2, \ldots, m\}$). Then, multiple versions of $\hat{X}$ are generated as $\{\hat{X}(\lambda_1), \hat{X}(\lambda_2), \ldots, \hat{X}(\lambda_m)\}$, where $\hat{X}(\lambda_k) = X + e(\lambda_k)$), each with increasing error variance (specifically, $e(\lambda_k)$ has variance $(1 + \lambda_k)\sigma_e^2$). In other words, the method simulates variables with increasingly larger measurement errors. Each $\hat{X}(\lambda_k)$ is associated with a set of coefficient estimates $\theta(\lambda_k)$. In the extrapolation step, a parametric model $\theta(\lambda)$ is estimated, which describes the relationship between the magnitude of the error and the coefficients. Then, extrapolating $\theta(\lambda)$ to $\theta(-1)$, one can approximate the coefficient estimates under zero measurement error (see Cook and Stefanski (1994) for more details). Figure 4.2 provides a graphical illustration of the SIMEX correction process. The parametric model $\theta(\lambda)$ may take several functional forms, including linear, quadratic, and nonlinear. Asymptotic methods have been proposed to estimate the standard errors for corrected coefficients following the application of the SIMEX method, including the delta (Carroll et al. 1996), jackknife (Stefanski and Cook 1995), and bootstrapping methods.

Figure 4.2. Graphical Illustration of the SIMEX Correction Process



Note. In the simulation step, four versions of X with increasing error are generated. Each corresponds to a set of parameter estimates, marked by points A, B, C, and D. In the extrapolation step, a parametric model (as shown by the dotted curve) is fitted, and extrapolated to the case where no error is present, marked by point E. The subplots on the right show the changes in regression line (obtained during the second-stage econometric estimation) during the error correction process.

The MC-SIMEX method, an extension of the SIMEX method, was introduced by Küchenhoff et al. (2006) to accommodate misclassification in *discrete* variables when the misclassification matrix is known or can be estimated. It involves the same two basic steps as SIMEX. In the simulation step, $\hat{X}(\lambda_k)$ is generated by adjusting the values of $\hat{X}$ based on the λ_k th power of the misclassification matrix (see Section 4.3.3 for the procedure of adjusting values of $\hat{X}$ based on a given misclassification matrix). In the extrapolation step, a parametric function $\theta(\lambda)$ is estimated and extrapolated to $\theta(-1)$, to approximate coefficient estimates under conditions of zero misclassification. Küchenhoff et al. (2007) proposed an asymptotic standard error estimation method for MC-SIMEX. Appendix 4.3 provides the pseudocode for implementing both SIMEX and MC-SIMEX methods.

For both classical measurement error and misclassification, SIMEX and MC-SIMEX can be directly applied, regardless of the second-stage model specifications. However, for non-classical measure error that contains systematic error, SIMEX correction is unlikely to be effective. Take the error structure $\hat{X} = a + bX + e$ and $b \neq 1$ as an example. Although SIMEX can eliminate estimation bias caused by the random component e, it cannot fix the bias from the systematic component. To overcome this challenge, we propose a data *pre-processing* step in addition to the original SIMEX procedure. Because we can observe both X and $\hat{X}$ in the labeled training data used to build first-stage data mining model, we can fit a linear regression of $\hat{X}$ on X to obtain estimations $\hat{a}$ and $\hat{b}$. We can then generate a new variable: $\widehat{X'} = (\hat{X} - \hat{a})/\hat{b}$, to reduce the non-classical error structure to the classical form $\widehat{X'} = X + e'$. From this relationship, we can calculate the modified error e' as the difference between $\widehat{X'}$ and X and its standard deviation as σ_e'. Then, we can apply the standard SIMEX correction procedure, using $\widehat{X'}$ as the (modified) variable with measurement error and σ_e' as (modified) error standard deviation.

4.4.3 Diagnosing error and evaluating correction efficacy

Before error correction, researchers should first assess three things. First, it is important to understand the error's functional form. If the measurement error contains a systematic component, it may require special error correction procedures such as the SIMEX procedure with data pre-processing we described in the previous section. Second, it is important to assess the severity of the bias in the second stage. While measurement error and misclassification may invalidate coefficient estimates and statistical inference, it is also possible to have trivial to minimal bias, which is of little concern. Third, it is important to evaluate the efficacy of the chosen error correction methods. For example, both regression calibration and SIMEX are general-purpose methods applicable to many circumstances. Researchers should carefully compare the relative efficacy of each correction procedure and choose the one that best fits their purposes.

In Table 4.4, we outline a basic procedure for diagnosing errors and choosing error correction methods. Following the procedure, researchers can use the labeled dataset from the first-stage data mining model to diagnose the functional form of the error, the severity level of bias, and the effectiveness of correction methods, because both the true values and model-predicted values of the variables are observed. Equipped with knowledge from the diagnostic procedure, researchers can proceed to actual analyses using the unlabeled dataset and apply the chosen error-correction method. The increase in sample size using unlabeled data may help identify desired effects with greater power and more precision.

4.4.4 Using SIMEX and MC-SIMEX for error correction

To demonstrate the effectiveness of SIMEX and MC-SIMEX, we applied them to the simulated data from Section 4.3. For each model specification, we ran either SIMEX (for continuous measurement error in X_1) or MC-SIMEX (for discrete misclassification in X_2) and reported the corrected coefficient estimates associated with the variables containing measurement error or misclassification. The efficacy of both methods depends on an accurate estimation of the extrapolation function $\theta(\lambda)$. Through experiments with simulated and actual data, researchers have identified the quadratic and nonlinear extrapolation functions to be effective for a large number of model specifications (Cook and Stefanski 1994; Küchenhoff et al. 2006). We used the quadratic extrapolation function for all of our simulations. Researchers should experiment with alternative extrapolation functions to determine the one best suited to their situation.

Table 4.4. Procedure for Diagnosing Error and Evaluating Correction Efficacy

Error Diagnostics (Steps 1-4):
Step 1: Conduct planned second-stage econometric analysis on the labeled dataset, using *true* labels.
Step 2: Conduct planned second-stage econometric analysis on the labeled dataset, using *model-predicted* labels.
Step 3: If error is continuous, use true labels and model-predicted labels to estimate error functional form.
Step 4: Compare estimates from Steps 1 and 2 to understand the impact of measurement error, including but not limited to (1) the degree of bias, (2) the direction of bias, (3) changes in statistical significance, and (4) changes in model fit. Use the estimate from Step 3 to understand the characteristics of the continuous error.
Correction Diagnostics (Steps 5-6):
Step 5: Apply candidate error-correction methods (e.g., SIMEX) on the second-stage econometric model. Use the estimate from Step 3, if warranted, to modify the error correction procedure(s) accordingly.
Step 6: Compare estimates from Steps 1, 2, and 4 to understand the efficacy of candidate error-correction methods, choose the most effective error-correction method for actual analysis.

Table 4.5a shows the correction results for classical measurement error models. Table 4.5b shows the results for non-classical measurement error models. For the systematic measurement error simulated in Scenario (3), we applied the SIMEX procedure with pre-processing. Table 4.5c shows results for our discrete misclassification models. In all the tables, the first two columns respectively contain coefficients without error and with error, denoted as b and b', and the third column contains the corrected estimation, denoted as b_c, obtained via SIMEX (Table 4.5ab) or MC-SIMEX (Table 4.5c). In Table 4.5b, we report corrected estimates, obtained from standard SIMEX procedure both without and with data pre-processing, denoted as b_c and b_{c_pre}. All

coefficients were statistically significant except those in parentheses.

Table 4.5a. SIMEX Correction for X_1 with Classical Measurement Error

	OLS			Logit			Probit			Poisson			Fixed-Effect		
	b	b'	b_c	b	b'	b_c	b	b'	b_c	b	b'	b_c	b	b'	b_c
$\sigma_e = 0.1$															
X_1	1.995	1.970	1.992	1.996	1.954	1.988	1.979	1.903	1.960	2.000	1.911	1.933	1.994	1.977	1.996
$\sigma_e = 0.3$															
X_1	1.995	1.833	1.998	1.996	1.748	1.999	1.979	1.567	1.878	2.000	1.760	1.977	1.944	1.824	1.985
$\sigma_e = 0.5$															
X_1	1.995	1.595	1.946	1.996	1.453	1.910	1.979	1.155	1.591	2.000	1.337	1.646	1.944	1.589	1.944

Table 4.5b. SIMEX Correction for X_1 with Non-Classical Measurement Error

	Scenario (1)			Scenario (2)			Scenario (3)			
	b	b'	b_c	b	b'	b_c	b	b'	b_c	b_{c_pre}
$\sigma_e = 0.1$										
X_1	1.995	1.884	1.901	1.995	1.979	1.999	1.995	3.830	3.985	1.997
$\sigma_e = 0.3$										
X_1	1.995	1.641	1.757	1.995	1.863	2.036	1.995	2.903	3.739	1.866
$\sigma_e = 0.5$										
X_1	1.995	1.412	1.634	1.995	1.667	2.057	1.995	1.941	2.922	1.456

Table 4.5c. MC-SIMEX Correction for X_2 with Misclassification

	OLS			Logit			Probit			Poisson			Fixed-Effect		
	b	b'	b_c	b	b'	b_c	b	b'	b_c	b	b'	b_c	b	b'	b_c
Scenario (1): $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.8$, and $M_{11} = 0.8$															
X_2	2.988	1.596	2.557	2.890	1.106	1.860	2.897	0.979	1.648	3.000	1.722	2.669	2.986	1.583	2.543
Scenario (2): $X_2 \sim Bernoulli(0.3)$, $M_{00} = 0.6$, and $M_{11} = 0.5$															
X_2	2.988	0.282	0.756	2.890	0.304	0.800	2.897	0.148	0.391	3.000	0.205	(0.565)	2.986	0.271	0.733
Scenario (3): $X_2 \sim Bernoulli(0.5)$, $M_{00} = 0.6$, and $M_{11} = 0.5$															
X_2	2.997	0.332	0.904	2.923	0.266	0.774	2.910	0.243	0.632	3.000	0.032	(0.065)	2.986	0.272	0.723

Based on Tables 4.5a and 4.5c, we can see that standard SIMEX and MC-SIMEX effectively reduced the bias in all regressions. In a number of cases, the correction procedure almost fully recovered the unbiased estimate. Even when misclassification was severe, such as in Scenarios (2) and (3) in Table 4.5c, MC-SIMEX enabled us to correct the coefficient of X_2 in the right direction, although the corrected coefficient was not statistically significant in the Poisson regressions. Our results from Table 4.5c also suggest that error correction methods have limited effectiveness when the performance of data mining model is poor. In these cases, researchers should focus on improving predictions first, and only deploy the correction methods as a secondary, remedial action.

Results in Table 4.5b show that SIMEX was also effective for non-classical measurement error. When measurement error was correlated with the true value of X_1, as in Scenario (1), SIMEX corrected the coefficient of X_1, although the correction was not as good as in the case of independent error. When measurement error was correlated with X_3, as in Scenario (2), SIMEX corrected the coefficients of both X_1 and X_3. When there was systematic error as in Scenario (3), SIMEX correction without pre-processing failed and actually exacerbated the bias in the coefficient of X_1, moving it further away from its true value. However, applying SIMEX after our proposed pre-processing successfully corrected the coefficient on X_1, for $\sigma_e = 0.1$ and $\sigma_e = 0.3$. For $\sigma_e = 0.5$, SIMEX with pre-processing also performed better than SIMEX without pre-processing, though it is worth noting that the corrected coefficient (1.456) was still further from the true value (1.995) than the original "biased" estimate (1.941). This marks an important situation under which the SIMEX method may be not only ineffective, but detrimental. When continuous measurement error contains both a systematic component and a random component with large variance, the two combined can result in a smaller "net" bias than each component alone. Under this special scenario, error correction is incapable of resolving the bias. Again, if the researcher first employs the diagnostic procedure outlined in Table 4.4, it would be possible to observe whether the chosen correction procedure is improving estimates, or in fact making matters worse.

Another important observation from Tables 4.5a-4.5c is that the effectiveness of SIMEX and MC-SIMEX corrections vary with (1) the amount of error and (2) the model specification. As the amount of measurement error or misclassification increases, the correction generally becomes less effective, i.e., the corrected coefficients shift further away from the true coefficients. Additionally, corrections for Linear and Logit models appear generally more effective than corrections for Probit and Poisson models.

To understand the effectiveness of SIMEX and MC-SIMEX corrections under a wider array of circumstances, we conducted additional, more comprehensive simulation studies. We extended the simulation studies described above by systematically varying the distributions and variances of the precisely measured covariates (i.e., X_2 and X_3 for simulations of measurement error in X_1; X_1 and X_3 for simulations of misclassification in X_2). Based on these additional simulations, we were able to further validate our aforementioned observations. First, SIMEX and MC-SIMEX are able to mitigate the biases in almost all cases. Importantly, as the amount of error increases, the magnitude of bias generally becomes larger, and the correction tends to become less effective. Second, corrections for Linear and Logit models appear to be more effective than corrections for Probit and Poisson models. Third, the effectiveness of corrections also depends on the distributions and variances of the error-free covariates. However, it is difficult to provide theoretical, *a priori* predictions about the correction's effectiveness for situations that we have not considered here. Accordingly, we would caution researchers to adopt the diagnostic procedure described in Table

4.4 in order to understand the nature of the error in their particular dataset, for their particular data mining model and regression specifications, and thereby assess the efficacy of any correction procedures in their unique empirical contexts.

4.5 Application to Field Data: Three Real-World Datasets

In this section, we apply SIMEX and MC-SIMEX methods to three real-world datasets. The three examples cover a variety of data types, model specifications, and research questions that are commonly seen in IS research. We use the first two examples to demonstrate the effectiveness of SIMEX and MC-SIMEX. We use the third example to illustrate a scenario under which the SIMEX correction is *not* effective, because of extremely poor performance of the predictive data mining model. In all three examples, we follow the diagnostic procedure outlined in Table 4.4, which helps to ascertain whether error correction is effective.

4.5.1 Review helpfulness on TripAdvisor.com

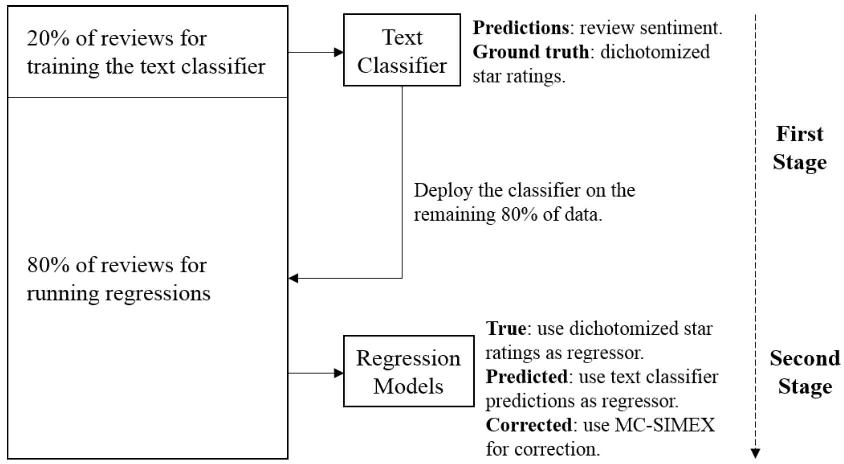
In the first example, we apply the MC-SIMEX method to a real-world dataset of online reviews from TripAdvisor.com. We examine the relationship between textual sentiment and perceived helpfulness, employing the two-stage approach of combining data mining and econometric modeling. We first built a textual classification model to predict the sentiment of written reviews as either positive or negative, and then estimated two econometric models controlling for several other factors. We drew on the star rating of a review as the ground truth for its sentiment.

There is an extensive body of literature on online reviews in the IS discipline. Researchers have investigated the effects of various review characteristics such as volume, valence, and reviewer identity on consumer behaviors (Mudambi and Schuff 2010; Forman et al. 2008; Dellarocas 2003). Some studies in this domain have also combined data mining with econometric analysis. Archak et al. (2011), for example, built a text classification model to identify product features from consumer product reviews, and then estimated the impact of specific product features on product sales.

TripAdvisor.com is a travel-related website that hosts consumer reviews of service providers. Users can post reviews about hotels, restaurants, or resorts. Reviewers provide an overall rating on a five-star scale and, optionally, ratings on separate dimensions of the consumption experience. For example, reviewers can rate a hotel based on its price, service, or overall quality. Readers of a review can indicate its “helpfulness” by casting a vote. As a prominent site for consumer reviews, TripAdvisor.com has been examined in several studies (e.g., Huang et al. 2016a; Huang et al. 2016b; Mayzlin et al. 2014).

We collected 11,953 English-language reviews for 234 randomly selected U.S. restaurants. For each review, we gathered data on its textual content, star rating, the number of helpful votes it received, whether the review contained a photo, and the number of reviews posted prior to the focal review, which indicated the review’s position in the sequence of all reviews for a restaurant. Using this dataset, we examined the impact of review sentiment on perceived helpfulness. Figure 4.3 shows the two-stage process.

Figure 4.3. Overview of the Two-Stage Process in Studying Review Helpfulness



To identify the sentiment of a review, we analyzed its textual content using natural language processing and textual classification techniques. In general, it is unnecessary to perform sentiment analysis when the star rating associated with a review is available, because a high rating often corresponds to positive sentiment and low rating corresponds to negative sentiment. However, many online venues host consumer opinions and word of mouth as text, without the benefit of numerical ratings (e.g., Godes and Mayzlin 2004). In such a setting, researchers typically hire a team of human coders to manually label the sentiment of a small, random sample of text from a large dataset. Using this labeled sample, one can then train a classifier and deploy it to classify the sentiment of the remaining unlabeled text. In this example, we treated the star rating of each review as the ground truth of its sentiment, for the purpose of training and evaluating a sentiment classifier that is based *only* on the textual content of reviews. Doing so allows us to quantify the misclassification and to illustrate the bias introduced in the second-stage econometric model due to error. If the reviewer gave a restaurant 3 or fewer stars, we coded the review as negative. If the reviewer gave 4 or 5 stars, we coded the review as positive. Using these criteria, 79% of the reviews in our sample were coded as positive and 21% were coded as negative, indicating a skewed distribution.

We followed standard practices in training the text classifier. First, we randomly selected 20% of the original sample (i.e., 2,391 reviews) as the labeled dataset for training and evaluating the performance of the model. Second, we followed standard natural language processing procedures (e.g., Jurafsky and Martin 2008) to convert each review into a word vector, in several steps. We transformed all text to lower-case, tokenized the text of each review into words, removed stop words, conducted stemming, and extracted bi-grams and tri-grams. We then applied the TF-IDF (term frequency-inverse document frequency) weighting scheme to rescale the word vector frequencies of occurrence (ibid). Third, we built a classifier using the linear Support Vector Machine (SVM) technique (Vapnik 1995), and evaluated the classifier using five-fold cross validation. Our classifier achieved 93.03% precision and 92.93% recall for the positive class, and 73.97% precision and 74.28% recall for the negative class. This performance corresponds to the following misclassification matrix: $(M_{00}, M_{10}, M_{01}, M_{11}) = (0.74, 0.07, 0.26, 0.93)$. Finally, we deployed the trained classifier on the remaining, unlabeled sample, i.e., on 9,562 reviews. In the end, every review in the unlabeled dataset had a predicted sentiment.

The dependent variable, *helpfulness*, was coded as a dummy variable indicating whether a review received any helpful votes. The independent variable, *sentiment*, was set to 1 if the review was positive and 0 if it was negative. We also included several control variables including: (1) *photo*, a dummy variable indicating whether the review had a photo or not; (2) *words*, the number of words in the review; and (3) *sequence*, the number of reviews posted about a restaurant before the focal review. We estimated two models, as illustrated in the equations below: a linear probability model (LPM) and a Logit model.

LPM: $helpfulness = \beta_0 + \beta_1 sentiment + \beta_2 photo + \beta_3 log(words) + \beta_4 sequence + \varepsilon$

Logit: $Logit(helpfulness) = \beta_0 + \beta_1 sentiment + \beta_2 photo + \beta_3 log(words) + \beta_4 sequence + \varepsilon$

Before carrying out the actual regression analysis and the MC-SIMEX correction, we followed the diagnostic procedure outlined in Table 4.4 by running the two regressions on our 20% labeled data (N = 2,391). We used five-fold cross validation to evaluate our first-stage SVM model. For each fold, we obtained the predicted sentiment label from the SVM model built off the other four folds. Our diagnostic analyses showed that misclassification in *sentiment* attenuated its effect on *helpfulness*, and that MC-SIMEX was effective in correcting the bias. We include these diagnostic results in Appendix 4.4. Table 4.6 shows our actual estimations, performed on the sample of 9,562 reviews.[30] For each model, we report three sets of results. The first column, labeled as "True", reports estimates obtained using the "true" values of the sentiment based on star ratings. The second column, labeled as "Predicted", reports estimates obtained using predicted sentiment from our text classifier. The third column, labeled as "Corrected", reports corrected estimates, by applying the MC-SIMEX method. We have provided the R code that was used to conduct the MC-SIMEX correction, in Appendix 4.5.

[30] Incorporating the labeled (i.e., ground truth) sample that was used to build the classification model may bias the misclassification matrix. However, in most research, because the labeled sample is usually a very small portion of the entire dataset, this bias in misclassification matrix generally will not invalidate the error-correction process.

Table 4.6. Regression Results and Corrections of the TripAdvisor.com Dataset (N = 9,562)

	LP Model			Logit Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1707***	0.1538***	0.1763***	-1.5750***	-1.6703***	-1.5644***
	(0.0105)	(0.0111)	(0.0166)	(0.0665)	(0.0715)	(0.0965)
Sentiment	**-0.0693*** **	**-0.0463*** **	**-0.0684*** **	**-0.4240*** **	**-0.2843*** **	**-0.3854*** **
	(0.0097)	**(0.0099)**	**(0.0161)**	**(0.0611)**	**(0.0629)**	**(0.0934)**
Photo	-0.0167*	-0.0174*	-0.0158*	-0.1149*	-0.1203*	-0.1100
	(0.0077)	(0.0077)	(0.0070)	(0.0578)	(0.0580)	(0.0568)
Words	0.7986***	0.7893***	0.7282***	4.3924***	4.3185***	3.9686***
	(0.0494)	(0.0510)	(0.0664)	(0.3011)	(0.3114)	(0.3715)
Sequence	-0.0010	-0.0049	-0.0040	0.0095	-0.0159	-0.0132
	(0.0166)	(0.0166)	(0.0174)	(0.1114)	(0.1114)	(0.1161)
Log Likelihood	-4408.24	-4422.64		-4470.65	-4483.95	
AIC	8828.5	8857.3		8951.3	8977.90	

Note. The MC-SIMEX method does not provide log likelihood or AIC statistics.

As shown in Table 4.6, sentiment was negatively associated with review helpfulness. Compared to positive reviews, negative reviews were more likely to receive helpful votes. In addition, reviews that contained photos were less likely to be perceived as helpful and longer reviews were more likely to be perceived as helpful. *Sequence* did not have a significant relationship with helpfulness. These findings were generally consistent with those of prior work (e.g., Mudambi and Schuff 2010; Yin et al. 2014), which indicates that our second-stage model specification was appropriate and valid.

Comparing the “Predicted” regressions with the “True” regressions show that the misclassification in the predicted sentiment considerably biased the estimation, as expected. The coefficient associated with *sentiment* in the “Predicted” estimation was only two-thirds the magnitude of the coefficient in the “True” estimation (i.e., the estimation based on dichotomized star ratings). Had we relied directly on the *sentiment* variable generated by the data mining model and ignored the misclassification, we would have greatly underestimated the magnitude of the effect of review sentiment on perceived helpfulness. The presence of misclassification in the *sentiment* variable also biased the other coefficient estimates, to various degrees. We also observed that the "Predicted" regressions exhibited worse model fit than the "True" regressions, assessed based on the log likelihood and AIC. Overall, the analysis proved the effectiveness of MC-SIMEX in correcting estimation bias from misclassification. This is particularly true in the LP Model, where MC-SIMEX almost perfectly recovered the true, unbiased coefficient estimate for *sentiment*.

To assess the impact of sample size on correction effectiveness, we repeated the above analyses for three random samples of 500, 2,000, and 5,000 observations. We observed three notable patterns. First, for each sample size, MC-SIMEX was able to mitigate the bias on *sentiment*. Second, as sample size increased from 500 to 5,000, the relative magnitude of bias decreased and the effectiveness of correction increased. This indicates that a sufficiently large sample is necessary to obtain both precise estimations and good correction outcomes. Third, further increasing sample size from 5,000 to 9,562 (i.e., the full sample) did not reduce the relative magnitude of bias, but did produce better corrected coefficients for *sentiment*. This suggests that having an increasingly larger sample does not eliminate bias, but generally does benefit error correction. The results of these additional analyses are included in Appendix 4.6.

4.5.2 User engagement on Facebook business pages

In the second example, we applied MC-SIMEX to another real-world dataset on user-generated posts on Facebook business pages. We examined the relationship between post sentiment and user engagement with a post, measured as the number of comments the post had received. Again, we first built a textual classification model to predict the sentiment of posts as either positive or negative, and then estimated two econometric specifications controlling for several other factors.

Facebook business pages is a feature that Facebook launched in 2007, which enable companies to interact with their customers on Facebook. Organizations use Facebook business pages primarily for marketing purposes by posting information about their products and services, offering coupons, as well as encouraging consumers to share positive word-of-mouth (Goh et al. 2013). Visitors of the business page can engage with both marketer-generated and user-generated posts through liking, commenting, or sharing (Goh et al. 2013). In this example, we examine how the sentiment of a user-generated post affects the number of comments it receives. We gathered 8,059 user-generated posts, all of which created in 2012, from the Facebook business pages of 39 consumer-oriented Fortune-500 companies such as airlines, banks, and retailers. For each post, we collected its textual content, poster ID, and the number of comments the post attracted. We hired Amazon Mechanical Turk workers to label the sentiment of the posts. We had five independent workers code each post, and used the majority (modal) rule to determine the sentiment. In total, 2,751 posts were labeled as positive, and 5,308 were labeled as negative. These manually labeled sentiments served as the ground truth for building the sentiment classifier, and for validating the performance of the MC-SIMEX correction procedure in our second stage estimation.

We built our sentiment classifier using a random sample of 10% of the labeled data (806 posts). We followed standard procedures in building the text classifier, as described previously in

Section 4.5.1. The classifier was built using the linear SVM technique, and evaluated using five-fold cross validation. Our classifier achieved 84.21% precision and 81.45% recall for the positive class (denoted as class 1), and 90.56% precision and 92.10% recall for the negative class (denoted as class 0). This performance corresponds to the following misclassification matrix: $(M_{00}, M_{10}, M_{01}, M_{11}) = (0.92, 0.19, 0.08, 0.81)$. We then deployed the trained classifier on the remaining 90% of our labeled sample (7,253 posts), and included the predicted sentiment in the second-stage econometric analysis.

In our econometric analysis, we examined the relationship between post sentiment and user engagement. The dependent variable, *comments*, was the number of comments each post received. The independent variable, *sentiment*, was coded as 1 if the post was positive and 0 if the post was negative. We controlled for several factors that may affect the level of engagement with a post including (1) *Log(Words)*, the log-transformed word count of each post; (2) *User Activeness*, the posting user's level of activeness, measured as the total number of posts that the user had created on the business page where the focal post appeared in 2012; (3) *Log(Popularity)*: the popularity level of the page on which the focal post was published, measured as the total number of user posts on the page in 2012; and (4) *Type*: the media type of the focal post assigned by Facebook such as link, photo, video, or status. Because our dependent variable is a count measure, we ran both linear regression and Poisson regression.

Next, we followed the procedure in Table 4.4 to conduct a diagnostic analysis before the second-stage estimation. We ran the proposed regressions with 10% of our labeled data (N = 806).

The diagnostic analysis showed that misclassification in *sentiment* attenuated its effect on *comments*, and MC-SIMEX was able to correct the bias (detailed results in Appendix 4.7). We then conducted the regression analysis with the remaining 7,253 posts. Table 4.7 shows the regression results and corrected coefficients.

Table 4.7. Regression Results (N = 7,253)

	OLS Model			Poisson Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	-2.9475***	-3.2001***	-2.9373***	-2.2148***	-2.3401***	-2.1591***
	(0.4804)	(0.4810)	(0.4878)	(0.1053)	(0.1053)	(0.2750)
Log(Words)	0.5840***	0.6177***	0.5470***	0.3412***	0.3551***	0.3049***
	(0.0408)	(0.0420)	(0.0469)	(0.0087)	(0.0089)	(0.0319)
Activeness	0.0303***	0.0302***	0.0305***	0.0108***	0.0109***	0.0110***
	(0.0045)	(0.0045)	(0.0045)	(0.0005)	(0.0005)	(0.0016)
Log(Popularity)	0.3107***	0.3151***	0.3195***	0.1753***	0.1782***	0.1824***
	(0.0470)	(0.0471)	(0.0471)	(0.0102)	(0.0102)	(0.0258)
Type = Link	-0.7544*	-0.7428*	-0.7504*	-0.5955***	-0.6051***	-0.6081*
	(0.3335)	(0.3343)	(0.3341)	(0.0994)	(0.0994)	(0.2383)
Type = Photo	0.0265	-0.1490	-0.0451	-0.0065	-0.1307	-0.0462
	(0.2679)	(0.2669)	(0.2684)	(0.0699)	(0.0693)	(0.1342)
Type = Video	-1.0167	-1.0116	-0.9993	-0.7607*	-0.8191**	-0.8086
	(0.9282)	(0.9304)	(0.9302)	(0.3017)	(0.3017)	(0.5908)
Sentiment	**-0.7356*** **	**-0.4789*** **	**-0.7132*** **	**-0.5724*** **	**-0.4047*** **	**-0.6307*** **
	(0.0987)	**(0.1041)**	**(0.1333)**	**(0.0251)**	**(0.0262)**	**(0.1002)**
Log Likelihood	-19559	-19576		-12701	-12859.5	
AIC	39133	39167		35653	35970	

Note. The MC-SIMEX method does not provide log likelihood or AIC statistics.

According to Table 4.7, positive posts received fewer comments than negative posts. Due to misclassification in *Sentiment*, all coefficient estimates were biased to various degrees and in different directions. The most important thing to note is that MC-SIMEX effectively mitigate the estimation biases in both OLS and Poisson models. For linear regression, MC-SIMEX almost fully recovered the unbiased coefficient of *Sentiment*. For Poisson regression, there was a slight overcorrection, i.e., the absolute value of the corrected coefficient of *Sentiment* was greater than its unbiased value, but the corrected estimate was still closer to the true value than the biased coefficient.

4.5.3 Campaign organizer age and crowdfunding outcomes

Our third and final example demonstrates the application of SIMEX correction to a real-world dataset of crowdfunding campaign outcomes from a leading reward-based crowdfunding website (Agrawal et al. 2014). We examined the relationship between the age of a fundraising campaign organizer and the amount of money he or she was able to raise. We collected campaign organizers' profile pictures and used a third-party face recognition service to infer the age of the persons in those pictures. The predicted age was not all accurate and contained measurement error. We used user's self-reported age, which we obtained from the platform operator, as the ground truth. We estimated a linear regression model controlling for several other factors. While we chose the first two examples to demonstrate the effectiveness of our proposed error correction methods, we chose this third example to show the limitation and boundary conditions of the methods, i.e., their effectiveness depends on a reasonable level of performance of the data mining models.

In recent years, crowdfunding has garnered a great deal of attention within the IS community (Burtch et al. 2013; 2015). On reward-based crowdfunding platforms like Kickstarter and Indiegogo, individuals can launch fundraising campaigns to raise money from the crowd to finance a project, a cause, or a venture. The money may be used to fund a new product or service or to support public goods and charitable endeavors. For each campaign, the organizer sets a fixed amount of money to be raised, and a fixed duration for the fundraising. A campaign is deemed a success if the fundraising goal is reached or surpassed within the specified duration. In this example, we examine the following research question: how does the age of a campaign organizer affect a campaign's fundraising success? Although age information is not directly available on the website, it can be inferred from organizers' profile pictures. We gathered information on 1,368 crowdfunding campaigns, each with a unique organizer who had uploaded a high-quality profile picture. For each campaign, we collected data on its beginning and end dates, the fundraising goal, the amount of money it collected by the end of the campaign, and whether it had been featured on the homepage of the crowdfunding website. We also had access to self-reported demographic information for each campaign organizer, including his or her gender and year of birth. We used the year of birth to calculate an organizer's actual age at the time of our data collection, which was used as the ground truth for the age variable. Next, we replicated the two-stage approach of combining data mining with econometric analysis.

In the first stage, we downloaded the profile pictures of the 1,368 campaign organizers. We used the Microsoft Face API,[31] a third-party face recognition service, to automatically infer the age of each organizer based on his or her profile picture. There were 63 profile pictures that contained more than one person. In those cases, we took the average of the predicted ages of all individuals appearing in the photo. Having both true and predicted ages for each organizer, we estimated the measurement error structure in the form of $\hat{X} = a + bX + e$, where $\hat{X}$ was the

[31] https://www.microsoft.com/cognitive-services/en-us/face-api

predicted age and X was the true age. We estimated the error structure on 30% of our sample, i.e., 410 randomly selected campaign organizers. This was done to mimic the reality that a researcher typically only has a small subsample of labeled data in practice. This analysis indicated that $\hat{a} = 18.78, \hat{b} = 0.36, SD(e) = 9.96$, which signaled very high levels of both systematic error and random error. Using error measures in data mining, such error corresponded to a MAE value of 10.58 and a RMSE value of 14.14 (in years). In the context of age recognition, aside from the inherent difficulty of estimating one's age based on a photo, there were other sources of measurement error such as cosmetic or photo-retouching effects, the use of someone else's photos, or the use of photos from a younger age.

In the second stage, we fit a linear regression to examine the relationship between organizer age and campaign outcomes. The dependent variable, *percent*, is the percentage of fundraising goal achieved by the end of a campaign. This variable can be greater than 1 if a campaign raised more than its fundraising goal. The independent variable, *age*, is either the true value or predicted value of the organizer's age. We included three control variables: (1) *gender*, representing the organizer's self-reported gender; (2) *featured*, a dummy variable indicating whether the campaign had been featured on the platform homepage at any point during the course of fundraising; and (3) *duration*, representing the number of days from the beginning to the end of the campaign. Given the existence of both systematic and random error components in measurement error, we used the SIMEX procedure with data pre-processing that we proposed in Section 4.4.

Before the second-stage regression analysis and the SIMEX correction, we followed the diagnostic procedure outlined in Table 4.4, by running the regression on the random subsample of 30% of our data, which was used to understand error functional form. Table 4.8a shows results of our diagnostic analysis, respectively the true coefficients, the predicted coefficients and two corrected estimations, from SIMEX procedures without and with pre-processing. Although we used SIMEX with pre-processing for error correction, we also included the corrected coefficients from SIMEX without pre-processing for comparison purposes. We observed that, while the true relationship between age and fundraising outcomes was negative, measurement error caused the sign to flip to positive. Furthermore, our results suggest that, without data pre-processing, SIMEX correction failed to mitigate the bias. Interestingly, even with pre-processing, SIMEX procedure was incapable of recovering the correct sign of the age variable. Diagnostic analysis suggests that the level of measurement error from the image classifier was too high to be corrected by SIMEX. We next carried out the actual second-stage analysis on the remainder of our data, and the results are shown in Table 4.8b.

We observed similar results from second-stage analysis. After controlling for other factors, age was negatively associated with the percentage of funding goal achieved. More specifically, an increase of 10 years in organizers' age can reduce fundraising by approximately 3.5%. Compared to younger people, older people were less likely to achieve their fundraising goals. Due to severe measurement error in age, the coefficient estimate of *age* became positive and insignificant when predicted age was included as the regressor. As such, measurement error introduced by the image classifier would, in this case, cause researchers to completely miss the effect of age. SIMEX correction without pre-processing further increased the bias, because it did not account for the systematic error component. Despite our best effort, even with pre-processing, SIMEX correction failed to recover the correct sign and significance of the coefficient on *age*, although it produced a coefficient that was closer to the unbiased value.

Table 4.8a. Regression Results for diagnostic (N = 410)

	Diagnostic Analysis			
	True	Predicted	Corrected (no pre-process)	Corrected (pre-process)
Intercept	0.4006*** (0.0805)	0.2826*** (0.0683)	0.2392* (0.1076)	0.2972*** (0.0486)
age	**-0.0020** **(0.0018)**	**0.0013** **(0.0019)**	**0.0027** **(0.0034)**	**0.0007** **(0.0010)**
gender = male	-0.1190** (0.0405)	-0.1190** (0.0406)	-0.1219** (0.0400)	-0.1207** (0.0399)
featured = yes	1.0676*** (0.1155)	1.0634*** (0.1159)	1.0571** (0.3532)	1.0608** (0.3536)
duration	-0.0005* (0.0002)	-0.0006** (0.0002)	-0.0006*** (0.0001)	-0.0006*** (0.0001)
R-squared	19.55%	19.4%		

Note. The SIMEX method does not provide R-squared statistics.

Table 4.8b. Regression Results for actual analysis (N = 1,368)

	Actual Analysis			
	True	Predicted	Corrected (no pre-process)	Corrected (pre-process)
Intercept	0.4017*** (0.0415)	0.2542*** (0.0338)	0.2442*** (0.0488)	0.2573*** (0.0243)
age	**-0.0035***** **(0.0009)**	**0.0003** **(0.0009)**	**0.0006** **(0.0015)**	**0.0002** **(0.0005)**
gender = male	-0.0598** (0.0205)	-0.0557** (0.0208)	-0.0567** (0.0211)	-0.0565** (0.0211)
featured = yes	0.8320*** (0.0535)	0.8291*** (0.0538)	0.8288*** (0.0538)	0.8288*** (0.0538)
duration	-0.0004*** (0.0001)	-0.0004*** (0.0001)	-0.0004*** (0.0001)	-0.0004*** (0.0001)
R-squared	17.24%	16.44%		

Note. The SIMEX method does not provide R-squared statistics.

Our third example demonstrates an important lesson and insight about both SIMEX method and error correction procedures in general. When data mining models perform poorly and generate predictions with severe error, it is very challenging to fully recover the sign and statistical significance of the true coefficient using any error correction method. In this example, our post hoc assessment suggests that the true age is only weakly correlated with machine-detected age ($\rho = 0.33$). In other words, we would like to note that error correction methods do not have unlimited capability of uncovering the correct signal from any amount of error. Knowing this, researchers should prioritize reducing error and improving performance in their first-stage data mining models, rather than rely primarily on error corrections.

4.6 Discussion and Conclusions

An increasing use of data mining and econometrics as a two-stage analysis process provides many new opportunities for IS research. In the first stage, a wide variety of data mining techniques equip researchers with the tools to classify unstructured data, such as text or images, and to gather information that is not directly observable, such as sentiment. The output of these models can be subsequently incorporated into the second-stage econometric estimations to test hypotheses and make inferences. This combined approach, however, has potential pitfalls. In particular, this practice introduces challenges to statistical inference because of the well-known issues of measurement error or misclassification, which may compromise researchers' ability to draw robust conclusions. As we have demonstrated both analytically and empirically, ignoring measurement error or misclassification is likely to severely bias econometric estimations. We have also shown that (1) even a relatively low level of measurement error or misclassification from data mining models can result in substantial biases in subsequent econometric estimations, and (2) the biases are harder to anticipate when the error structures or econometric models grow more complex. This issue is particularly concerning, given the increasing focus on the magnitude of coefficients (i.e., the economic significance) in empirical studies, over and above mere statistical significance (Lin et al. 2013).

Fortunately, standard practices in data mining involve the evaluation of model performance using test datasets and provide established ways to quantify measurement error or misclassification. With this information, we can take actions to mitigate biases in the second-stage estimation. In this essay, we reviewed several error-correction methods and focused on two simulation-based methods, SIMEX and MC-SIMEX, as promising remedies to correct for biases from measurement error and misclassification. We illustrated their effectiveness using both comprehensive simulations and three real-world datasets. In most cases, biases were reduced and the corrected coefficients were closer to the true values. In some cases, the corrected coefficients almost perfectly recovered their true values. We also identified two situations under which the effectiveness of error correction methods may be either limited or unnecessary. First, when the level of measurement error or misclassification is very high, SIMEX or MC-SIMEX are not powerful enough to correct the bias. Solely relying on these methods could lead researchers to draw incorrect conclusions. Second, when continuous measurement error contains both a systematic component and a random component with relatively large variability, it can sometimes result in little bias in coefficient estimation. Using SIMEX in this situation could therefore be unnecessary or even detrimental. Finally, note that error correction methods cannot account for biases caused by misspecification in the second-stage econometric model. For example, we simulated scenarios where the regression model had omitted variable bias, besides measurement error. The SIMEX method was able to correct for the bias due to the measurement error, but it had no way of identifying the existence of omitted variable bias.

In addition to causing biases in regression coefficient estimates, measurement error and misclassification in independent variables can affect several other important aspects of econometric analysis, including confidence interval estimation, goodness-of-fit calculation, and hypothesis testing. For a linear regression where one of the independent variables contains *classical* measurement error, several asymptotic results are known in literature. First, estimation of the error variance σ_ε^2 will be inconsistent (Wansbeek and Meijer 2000). More specifically, the sample estimate s_ε^2 will exceed the true value σ_ε^2 in the limit. As a result, standard error for each regressor will also be overestimated in the limit. Consequently, the corresponding confidence interval will be wider than it should be, and the corresponding p-value will be larger than it should be. In general, classical measurement error in linear regressions makes OLS estimators more conservative. Second, as a direct ramification of the overestimation of error variance, R^2 is biased toward zero, indicating worse model fit. Third, the reliability of hypothesis testing may become questionable. For example, the commonly used *F*-statistic is biased toward zero, which means the null hypothesis that every coefficient is zero is not rejected often enough. In the case of misclassification, although limited theoretical results are available in the literature, our simulations showed that the presence of misclassification can inflate estimation of model error variance and can result in decreased model fit. Although we primarily focus on demonstrating and correcting biases in coefficient estimates in this essay, we believe that readers should be aware of the aforementioned other consequences of error.

This essay highlights both the opportunities and potential pitfalls of combining data mining and econometric modeling. Given the growing prevalence of the integrated approach, we hope to raise awareness of the fact that failing to account for measurement error or misclassification, which arises from the data mining process, could result in misleading findings. We chose SIMEX and MC-SIMEX as exemplary error-correction methods because they are easy to parameterize by using performance metrics from the data mining process and because they can be applied to a variety of econometric models. However, we do not claim that these two methods are superior to other error-correction methods in all situations. Instead, we acknowledge there are situations where other methods may be more appropriate. For example, the regression calibration method has been shown to produce consistent estimates for linear models (Carroll et al. 2006), and the instrumental variable approach can be used when valid instruments are available. We encourage researchers to evaluate and adopt error correction methods on a case-by-case basis, depending on the nature of their data and research setting. We provide a diagnostic procedure to help researchers assess and deal with measurement error in their research practice. We propose that researchers use labeled dataset from first-stage data mining to diagnose the structure of the error, the severity of resulting bias, and the effectiveness of available correction methods. Conducting these diagnostic analyses before applying the error correction procedure can help the researcher fully understand and address the issue.

Our essay provides a first step toward addressing the challenges with measurement error from combining data mining techniques with econometric analysis. There are several promising avenues for future work. The first future direction is to continue improving existing error-correction methods. When applying the SIMEX and MC-SIMEX methods, we occasionally observed cases in which the coefficients of precisely measured (error-free) covariates were slightly over-corrected or shifted in the opposite direction. Although the mathematical underpinnings of the correction methods in no way would suggest that this result is a systematic or asymptotic property (but rather is a finite sample property), researchers should be aware of this potential issue. Future research should continue to improve the stability and robustness of the SIMEX and MC-SIMEX methods. Second, there are challenging scenarios when several variables, potentially of different types, are simultaneously measured with error, or when the measurement error takes complicated forms. Current error correction methods may not be capable of mitigating biases in these challenging cases, which calls for more novel and powerful new methods. Third, researchers can seek to develop novel approaches to combine predictive data mining with econometric analysis that avoid the peril of measurement error. Through this essay, we hope to raise awareness of these methodological challenges and opportunities and help IS scholars to better sharpen our collective toolkit and harness the power of data mining methods in empirical research.

Chapter 5. Concluding Remarks

My dissertation examines UGC and the associated user engagement behaviors on social media brand pages such as Facebook business pages. These business pages empower individual users to engage with the focal businesses and with each other. Despite its recognized importance, user engagement toward UGC on these business pages has been under-explored. My research fills in this gap. In three essays, I respectively discuss (1) valence and content characteristics of user-generated posts on Facebook business pages, and the effects of valence/content antecedents on two distinct engagement behaviors (i.e., liking and commenting); (2) the interplay among multiple engagement features, by which engagement behaviors are shaped; and (3) a methodological challenge in drawing statistical inference from econometric modeling that incorporates independent variables generated from data mining, which has been largely ignored in prior literature that often used this combined methodology to study textual content in online contexts (e.g., UGC on social media platforms).

My dissertation research contributes new theoretical understanding and empirical evidence to the Information Systems literature. In the first essay, the results highlight user posts on Facebook business pages as a new form of online UGC that is qualitatively distinct from the extensively studied online consumer reviews, in terms of both valence distribution and salient content categories. Users engage with user posts through engagement features such as Likes and Comments. Notably, different forms of engagement are driven by different engagement goals and motivations, and are affected by different, sometimes opposing, factors.

While the first essay establishes the unique characteristics of engagement behaviors towards user posts on Facebook business pages, the second essay demonstrates that the usage of different engagement features is not independent. Instead, user posts that have received engagement may be perceived to be more worthy of attention, and continue to attract even more engagement. At the same time, the introduction of a new engagement feature may "widen the gap", leading the engaging content to receive more engagement, and the unengaging content to receive less engagement, than what they would have received before the new feature introduction.

Finally, the third essay advances the methodological rigor of studying UGC or other unstructured data in various contexts, and helps researchers to integrate data mining methods and econometric methods in a robust manner. Directly adding predictions from a data mining model into an econometric model as independent regressor would bring measurement error and cause estimation biases. Fortunately, performance metrics obtained from data mining model evaluation readily quantify the amount of measurement error, which provides a unique opportunity to correct for estimation biases, using statistical methods such as simulation-extrapolation. To guide this error-correction practice, we encourage researchers to adopt the proposed diagnostic procedure. The third essay constitutes an important methodological contribution to the Information Systems as well as other domains.

Findings from my dissertation research can also generate actionable practical implications, both for social media platforms and for the businesses that use social media to build their digital presence. For social media platforms, it is useful to be cognizant about the behavioral consequences of key design features, such as the user post feature and the Reactions feature, and to understand how users actually use the features. This can help the platforms design meaningful features and properly measure their efficacy. For large companies in consumer-facing industries, a deep understanding of user posts and the engagement behaviors can facilitate better management of their Facebook business pages. Given a large number of negative posts, companies need dedicated social media strategies to manage unfavorable user voices, which should take into account the unique user behaviors on social media. For example, companies may want to prioritize in addressing and responding to user content that has received engagement from other users, because such content likely represents issues that many users care about, and can create far-reaching impact on the users involved. Further exploration of these practical implications and their impact on businesses and consumers also represents important future research directions.

Bibliography

Agarwal, R., & Dhar, V. (2014). Editorial—big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25*(3), 443-448.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Aggarwal, R., Gopal, R., Gupta, A., & Singh, H. (2012). Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Information Systems Research*, *23*(3), 976-992.

Agrawal, A., Catalini, C., & Goldfarb, A. (2014). *Some Simple Economics of Crowdfunding*. In J. Lerner & S. Stern (Eds.), Innovation Policy and the Economy (1st ed., Vol. 14, pp. 63–97). Chicago, IL: University of Chicago Press.

Allison, P. D., & Waterman, R. P. (2002). Fixed–effects negative binomial regression models. *Sociological Methodology*, 32(1), 247-265.

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, *1*(1), 5-17.

Anderson, M., Sims, J., Price, J., & Brusa, J. (2011). Turning "Like" to "Buy" social media emerges as a commerce channel. *Booz & Company Inc*.

Angrist, J. D., & Pischke, J. S. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special issue-social media and business transformation: A framework for research. *Information Systems Research*, 24(1), 3-13.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8), 1485-1509.

Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., & Wang, X. (2006, April). Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 959-968). ACM.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371-1391.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of general psychology*, *5*(4).

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586-607.

Bhattacherjee, A. (2001). Understanding information systems continuance: an expectation-confirmation model. *MIS Quarterly*, *25*(3), 351-370.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Brodie, R. J., Hollebeek, L. D., Juric, B., & Ilic, A. (2011). Customer engagement: conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research*, 1-20.

Brodie, R. J., Ilic, A., Juric, B., Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. *Journal of Business Research*, 66: 105-114.

Buonaccorsi, J. P., Laake, P., & Veierød, M. B. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, *61*(3), 831-836.

Burke, M., & Kraut, R. E. (2014). Growing closer on Facebook: changes in tie strength through social network site use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 4187-4196.

Burtch, G., Ghose, A., & Wattal, S. (2013). An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research*, 24(3), 499-519.

Burtch, G., Ghose, A., & Wattal, S. (2015). The hidden cost of accommodating crowdfunder privacy preferences: a randomized field experiment. *Management Science*, 61(5), 949-962.

Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, *91*(433), 242-250.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Cavusoglu, H., Phan, T. Q., Cavusoglu, H., & Airoldi, E. M. (2016). Assessing the impact of granular privacy controls on content sharing and disclosure on Facebook. *Information Systems Research*, *27*(4), 848-879.

Chan, J., & Wang, J. (2014). Hiring Biases in Online Labor Markets: The Case of Gender Stereotyping. *Proceedings of the International Conference on Information Systems (ICIS)*, Auckland, New Zealand.

Chan, K. (2009). I like this. Retrieved from https://www.facebook.com/notes/facebook/i-like-this/53024537130/.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165-1188.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, 54(3), 477-491.

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

Cho, Y., Im, I., Hiltz, R., & Fjermestad, J. (2002). An analysis of online customer complaints: implications for web complaint management. *Proceedings of the 35th HICSS (pp. 2308-2317).*

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*(428), 1314-1328.

Corstjens, M., & Umblijs, A. (2012). The Power of Evil: The Damage of Negative Social Media Strongly Outweigh Positive Contributions. *Journal of advertising research*, *52*(4), 433-449.

Cvijikj, I. P., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining*, 3(4), 843-861.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375-1388.

De Veirman, M., Cauberghe, V., Hudders, L., & De Pelsmacker, P. (2017). Consumers' Motivations for Lurking and Posting in Brand Communities on Social Networking Sites. *Digital Advertising: Theory and Research*, 207.

De Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, *26*(2), 83-91.

Dellarocas, C. (2003). The digitization of word-of-mouth: promise and challenges of online feedback. *Management Science*, 49(10), 1407–1424.

Dellarocas, C., Gao, G., & Narayan, R. (2010). Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems*, *27*(2), 127-158.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23-45.

Dessart, L., Veloutsou, C., & Morgan-Thomas, A. (2015). Consumer engagement in online brand communities: a social media perspective. *Journal of Product & Brand Management*, *24*(1), 28-42.

Dholakia, U. M., & Durham, E. (2010). One café chain's Facebook experiment. *Harvard Business Review*, 88(3), 26.

Dillon, A., & Morris, M. G. (1996). User acceptance of new information technology: theories and models. In *Annual review of information science and technology*. Medford, NJ: Information Today.

Dong, X., Wang, T., & Benbasat, I. (2016). IT Affordances in Online Social Commerce: Conceptualization Validation and Scale Development. In *Proceedings of Twenty-second Americas Conference on Information Systems*, San Diego.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems*, 45(4), 1007-1016.

Facebook Newsroom. (2016). Retrieved from http://newsroom.fb.com/news/2016/02/reactions-now-available-globally/.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291-313.

Gao, G., Greenwood, B. N., Agarwal, R., & McCullough, J. S. (2015). Vocal minority and silent majority: how do online ratings reflect population perceptions of quality, *MIS Quarterly, 39*(3), 565-590.

Geraci, M. 2016. "Qtools: A Collection of Models and Tools for Quantile Inference," *R JOURNAL* (82), pp. 117-138.

Gerlitz, C., & Helmond, A. (2011). Hit, link, like and share. Organising the social and the fabric of the web. In *Digital Methods Winter Conference Proceedings,* 1-29.

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society* (15:8), 1348-1365.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498-1512.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3), 493-520.

Glaser, B.G., & Strauss, A.K. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine de Gruyter.

Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. *Contemporary Mathematics*, 112, 99-114.

Godes, D. (2011). Commentary-Invited Comment on "Opinion Leadership and Social Contagion in New Product Diffusion". *Marketing Science*, *30*(2), 224-229.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545-560.

Godes, D., & Mayzlin, D. (2009). Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science*, *28*(4), 721-739.

Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, *24*(1), 88-107.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*(2), 348.

Gu, B., Konana, P., Raghunathan, R., & Chen, H. M. (2014). The Allure of Homophily in Social Media: Evidence from Investor Responses on Virtual Communities. *Information Systems Research*, *25*(3), 604-617.

Gu, B., Konana, P., Rajagopalan, B., & Chen, H. M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research*, *18*(1), 68-85.

Gummerus, J., Liljander, V., Weman, E., & Pihlström, M. (2012). Customer engagement in a Facebook brand community. *Management Research Review*, *35*(9), 857-877.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

Hardin, J. W., Schmiediche, H., & Carroll, R. J. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata Journal*, 3(4), 373-385.

Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an

application to the patents-R&D relationship. *NBER*.
Hennig-Thurau, T., Gwinner, K. P., Walsh, G., Gremler, D. D. (2004). Electronic word-of-mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1): 38-52.
Hennig-Thurau, T., Walsh, G., & Walsh, G. (2003). Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the Internet. *International journal of electronic commerce*, *8*(2), 51-74.
Henning-Thurau, T., Gwinner, K. P., Walsh, G., Gremler, D. D. (2004). Electronic word-of-mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1): 38-52.
Hilbe, Joseph M. (2011), Negative Binomial Regression. 2nd edition. Cambridge University Press.
Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing?. *MIT Sloan Management Review*, 52(1), 41.
Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229-247.
Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144-147.
Huang, N., Burtch, G., Hong, Y., & Polman, E. (2016a). Effects of Multiple Psychological Distances on Construal Level: A Field Study of Online Reviews. *Journal of Consumer Psychology*, 26(4), 474-482.
Huang, N., Hong, Y., & Burtch, G. (2016b). Social Network Integration and User Content Generation: Evidence from Natural Experiments. *MIS Quarterly*, Forthcoming.
Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1-24.
Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*(2), 195-212.
Iyer, G., & Katona, Z. (2015). Competing for attention in social communication markets. *Management Science*, *62*(8), 2304-2320.
Jelveh, Z., Kogut, B., & Naidu, S. (2014). Political language in economics. *Available at SSRN 2535453*.
Johnson, S. L., Safadi, H., & Faraj, S. (2015). The Emergence of Online Community Leadership. *Information Systems Research*, 26(1), 165-187.
Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer‐Mediated Communication*, *11*(3), 723-747.
Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing*. Prentice Hall.
Khobzi, H., Lau, R. Y. K., & Cheung, T. C. H. (2017). Consumers' Sentiments and Popularity of Brand Posts in Social Media: The Moderating Role of Up-votes. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
Kim, Y. H., Kim, D. J., & Wachter, K. (2013). A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision Support Systems*, 56, 361-370.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. Retrieved from http://j. mp/1sexgVw.

Kiron, D., Palmer, D., Phillips, A. N., & Berkman, R. (2013). Social business: Shifting out of first gear. *MIT Sloan Management Review Research Report*.

Kraut, R. E., Rice, R. E., Cool, C., & Fish, R. S. (1998). Varieties of social influence: The role of utility and norms in the success of a new communication medium. *Organization Science, 9*(4), 437-453.

Küchenhoff, H., Lederer, W., & Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics & Data Analysis*, *51*(12), 6197-6211.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*(1), 85-96.

Kumar, N., Qiu, L., & Kumar, S. (2017). Exit, Voice, and Response in Digital Platforms: An Empirical Investigation of Online Management Response Strategies. *Information Systems Research*, forthcoming.

Lee, D., Hosanagar, K., & Nair, H. S. (2017). Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science, forthcoming*.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906-917.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3), 74-89.

Liu, Y., Chen, R., Chen, Y., Mei, Q., & Salib, S. (2012). I loan because...: Understanding motivations for pro-social lending. In *Proceedings of the 5th ACM international conference on Web search and data mining* (pp. 503-512).

Lu, X., Ba, S., Huang, L., & Feng, Y. (2013). Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Information Systems Research*, 24(3), 596-612.

Lu, Y., Jerath, K., & Singh, P. V. (2013). The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Management Science*, *59*(8), 1783-1799.

Ludford, P. J., Cosley, D., Frankowski, D., & Terveen, L. (2004). Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 631-638, ACM.

Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research*, *24*(1), 146-163.

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 Science*, *34*(5), 627-645.

Majchrzak, A., Faraj, S., Kane, G. C., & Azad, B. (2013). The contradictory influence of social media affordances on online communal knowledge sharing. *Journal of Computer‐Mediated Communication*, *19*(1), 38-55.

Malhotra, A., Malhotra, C. K., & See, A. (2013). How to create brand engagement on Facebook. *MIT Sloan Management Review*, *54*(2), 18.

Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: introducing the social CRM house. *Journal of Interactive Marketing*, *27*(4), 270-280.

Mangold, W., Miller, F., & Brockway, G. R. (1999). Word-of-mouth communication in the service marketplace. *Journal of Services Marketing*, 13(1), 73-89.

Maslowska, E., Malthouse, E. C., & Collinger, T. (2016). The customer engagement ecosystem. *Journal of Marketing Management*, *32*(5-6), 469-501.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *The American Economic Review*, 104(8), 2421–2455.

McAlexander, J. H., Schouten, J. W., & Koenig, H. F. (2002). Building brand community. *Journal of Marketing*, *66*(1), 38-54.

Miller, A. R., & Tucker, C. (2013). Active social media management: the case of health care. *Information Systems Research*, *24*(1), 52-70.

Mohr, L. A., & Webb, D. J. (2005). The effects of corporate social responsibility and price on consumer responses. *Journal of Consumer Affairs*, 39(1), 121-147.

Moreno, A., & Terwiesch, C. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research*, *25*(4), 865-886.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34(1), 185-200.

Muniz Jr, A. M., & O'guinn, T. C. (2001). Brand community. *Journal of Consumer Research*, 27(4), 412-432.

Muntinga, D. G., Moorman, M., Verlegh, P. W., & Smit, E. G. (2017). Who creates brand-related content, and why? The interplay of consumer characteristics and motivations. *Digital Advertising: Theory and Research*, 259.

Nan, N., & Lu, Y. (2014). Harnessing the Power of Self-Organization in an Online Community During Organizational Crisis. *MIS Quarterly, 38*(4), 1135-1157.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1-32.

Nimako, S. G., & Mensah, A. F. (2012). Motivation for customer complaining and non-complaining behaviour towards mobile telecommunication services. *Asian Journal of Business Management*, *4*(3), 310-320.

Oestreicher-Singer, G., & Sundararajan, A. (2012). The visible hand? Demand effects of recommendation networks in electronic markets. *Management science*, *58*(11), 1963-1981.

Oestreicher-Singer, G., & Zalmanson, L. (2013). Content or community? A digital business strategy for content providers in the social age. *MIS Quarterly*, 37(2), 591-616.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

Patterson, P., Yu, T., & De Ruyter, K. (2006). Understanding customer engagement in services. *Proceedings of ANZMAC 2006 conference*, Brisbane (pp. 4-6).

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, *10*(6), 601-626.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, *1*(1), 13-32.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-164.

Rhue, L. (2015). Who Gets Started on Kickstarter? Demographic Variations in Fundraising Success. *Proceedings of the International Conference on Information Systems (ICIS)*, Fort Worth, Texas.

Richins, M. L. (1983). Negative word-of-mouth by dissatisfied consumers: a pilot study. *The Journal of Marketing*, 68-78.

Richins, M. L., & Root-Shaffer, T. (1988). The role of involvement and opinion leadership in consumer Word-of-Mouth: An implicit model made explicit. *Advances in Consumer Research*, 15(1).

Rishika, R., Kumar, A., Janakiraman, R., & Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information systems research*, *24*(1), 108-127.

Rosenbaum, P. R. (2002). *Design of Observational Studies*, New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: the role of message content and style. *Journal of Consumer Behavior*, 11(3), 234-243.

Schöndienst, V., Kulzer, F., and Günther, O. (2012). Like versus dislike: How Facebook's like-button influences people's perception of product and service quality. In *Proceedings of the Thirty Third International Conference on Information Systems*, Orlando.

Scissors, L., Burke, M., & Wengrovitz, S. (2016). What's in a Like?: Attitudes and behaviors around receiving Likes on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1499-1508). ACM.

Shen, W., Hu, Y. J., & Ulmer, J. R. (2015). Competing for Attention: An Empirical Study of Online Reviewers' Strategic Behavior. *MIS Quarterly*, *39*(3), 683-696.

Shevlin, R. (2007). The value of customer engagement. Extracted from http://marketingroi.wordpress.com/ 2007/11/30/the-value-of-customer-engagement/

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

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?. *Journal of Interactive Marketing*, 26(2), 102-113.

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1.

Smock, A. D., Ellison, N. B., Lampe, C., & Wohn, D. Y. (2011). Facebook as a toolkit: A uses and gratification approach to unbundling feature use. *Computers in Human Behavior*, *27*(6), 2322-

2329.

Stefanski, L. A., & Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90 (432), 1247-1256.

Sundaram, D. S., Mitra, K., & Webster, C. (1998). Word-of-Mouth Communications: A Motivational Analysis. *Advances in consumer research*, *25*(1).

Sundararajan, A., Provost, F., Oestreicher-Singer, G., & Aral, S. (2013). Research commentary—information in digital, economic, and social networks. *Information Systems Research, 24*(4), 883-905.

Swani, K., Milne, G., & P. Brown, B. (2013). Spreading the word through likes on Facebook: Evaluating the message strategy effectiveness of Fortune 500 companies. *Journal of Research in Interactive Marketing, 7*(4), 269-294.

Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198-215.

Van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer engagement behavior: theoretical foundations and research directions. *Journal of Service Research*, *13*(3), 253-266.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Varian, H. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*, 28(2), 3-28.

Vivek, S. D., Beatty, S. E., Morgan, R. M. (2012). Consumer engagement: Exploring customer relationships beyond purchase. *Marketing Theory and Practice*, 20(2), 122-146.

Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing, 26*(2), 83-91.

Wang, T., Kannan, K. N., & Ulmer, J. R. (2013). The association between the disclosure and the realization of information security risk factors. *Information Systems Research*, 24(2), 201-218.

Wang, X., Butler, B. S., & Ren, Y. (2013). The impact of membership overlap on growth: An ecological competition view of online groups. *Organization Science*, *24*(2), 414-431.

Wansbeek, T., & Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics*. Elsevier.

Whittaker, S., Terveen, L., Hill, W., & Cherny, L. (2003). The dynamics of mass interaction. *From Usenet to CoWebs* (pp. 79-91). Springer London.

Wojnicki, A. C., & Godes, D. (2011). Signaling success: Strategically-positive word of mouth. *Working Paper,* Rotman School of Management, University of Toronto.

Wu, L. (2013). Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Information Systems Research*, 24(1), 30-51.

Yang, M., Ren, Y., & Adomavicius, G. (2014). Understanding word of mouth and customer engagement on Facebook Business Pages. In *Conference on Information Systems and Technology,* San Francisco, CA.

Yin, D., Bond, S., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539-560.

Zhang, S., Lee, D., Singh, P., Srinivasan, K. (2016) How Much Is An Image Worth? An Empirical Analysis of Property's Image Aesthetic Quality on Demand at AirBNB. *International Conference in Information Systems (ICIS), Dublin, Ireland*.

Zhang, X., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review*, *101*(4), 1601-1615.

Zhang, Y., Feick, L., & Mittal, V. (2014). How males and females differ in their likelihood of transmitting negative word of mouth. *Journal of Consumer Research*, *40*.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133-148.

Zhu, H., Kraut, R. E., Wang, Y. C., & Kittur, A. (2011). Identifying shared leadership in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3431-3434). ACM.

Zhu, H., Kraut, R., & Kittur, A. (2012). Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 407-416). ACM.

Appendices

Appendix 2.1: A sample task on MTurk to label post valence

Facebook Posts Sentiment

In this HIT, you will read a post made by a user on a company's Facebook page. Please carefully read the post and then choose its overall sentiment as positive, negative, or neutral. If a post has both positive and negative contents, you should choose the sentiment that is most applicable to the post as a whole. Please make your decision solely based on the text listed below.

Here is a post from the Facebook page of **Southwest Airlines**:

" Thanks for singing happy birthday to me on my flight from Ontario to Las Vegas! "

What is the overall sentiment of this post? Please select one of the following options.

- ○ Positive
- ○ Negative
- ○ Neutral

Submit

Appendix 2.2: A sample task on MTurk to label post content and the corresponding instruction

Instructions of How to Perform the HIT

Below we describe the categories. Please read the descriptions carefully.

- Positive Testimonial and Appreciation: the post includes a positive testimonial or a form of appreciation for the company (e.g., saying how wonderful the company is or how much the user loves it, thanking the company)
- Complaint about Product and Service Quality: the post includes a complaint about product and service quality of the company (e.g., poor quality products or bad services)
- Complaint about Money Issues: the post includes a complaint about money issues with the company (e.g., hefty fees or high prices)
- Other Complaint about the Company: the post includes a complaint about the company but it's NOT about product/service quality or money issues. Instead, for example, it may be a complaint about the company's standing on social or environmental issues such as labor, human rights, social equality, or pollution
- Customer Question: the post includes a question directed at the company (e.g., inquiry about its products)
- Customer Suggestion: the post includes a customer suggestion to the company (e.g., recommendation of new products and service to offer)
- Irrelevant Message: the post has nothing to do with the company on whose page this post appears. It may be user self promotion, promotional links, adult content, etc.
- None of the above: the post doesn't belong to any of the above categories

Facebook Posts Categorization

Below you will see a post from **US Airways's** fan page.

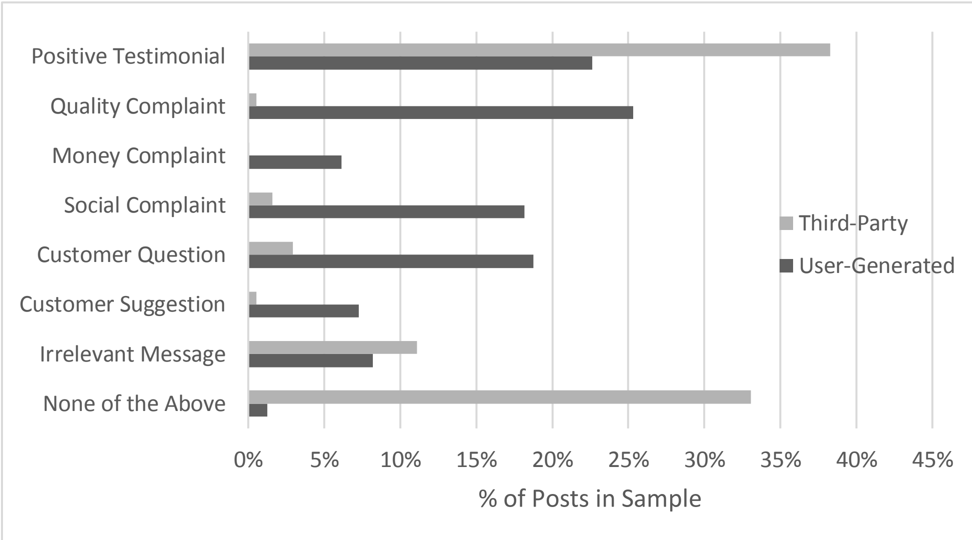
Can I ask a question here?? How many bottles of liquor I can bring into the US from Mexico??

Please read the post carefully and check all categories that apply. If no categories apply, check "none of the above" and enter a new category to describe the post.

- ☐ Positive Testimonial and Appreciation
- ☐ Complaint about Product and Service Quality
- ☐ Complaint about Money Issues
- ☐ Other Complaint about the Company
- ☐ Customer Question
- ☐ Customer Suggestion
- ☐ Irrelevant Message that has nothing to do with the Comany
- ☐ None of the above (Type your own category here) ____________

Submit

Appendix 2.3: Content category comparison between user-generated posts and third-party posts



Note. From the initial sample of 12,000 posts, we first removed 174 posts that had fewer than 2 words or fewer than 6 characters. After that, there were 10,705 user-generated posts and 1,121 third-party posts, which had meaningful content category labels. A post can belong to multiple categories, and we classified a post in a category only if three or more turkers agreed on that category.

Appendix 2.4: Regression results using conditional fixed effects negative binomial models

	Likes	Likes	Comments	Comments
Log(Word Count)	0.0982***	0.1448***	0.2602***	0.2235***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0107***	0.0083**	-0.0080**	-0.0047
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3761***	-0.3631***	-0.2441***	-0.2178***
	(0.0572)	(0.0589)	(0.0639)	(0.0636)
Log(Post-Level UGC)	0.2195***	0.1366***	-0.2408***	-0.1905***
	(0.0127)	(0.0133)	(0.0147)	(0.0155)
Log(Post-Level MGC)	-0.1134***	-0.0865***	0.0811***	0.0687***
	(0.0228)	(0.0225)	(0.0168)	(0.0169)
LexisNexis_1	0.0132***	0.0112***	0.0005	0.0022
	(0.0023)	(0.0023)	(0.0023)	(0.0022)
User Activeness	0.0042***	0.0027**	0.0023*	0.0030**
	(0.0010)	(0.0010)	(0.0011)	(0.0010)
Positive Valence	0.5385***		-0.3539***	
	(0.0490)		(0.0378)	
Negative Valence	0.7733***		0.0490	
	(0.0473)		(0.0337)	
Positive Testimonial		0.2452***		-0.2098***
		(0.0509)		(0.0429)
Quality Complaint		0.0984*		0.2715***
		(0.0469)		(0.0365)
Money Complaint		0.1264		0.1300**
		(0.0667)		(0.0467)
Social Complaint		0.9280***		-0.2088***
		(0.0500)		(0.0518)
Customer Question		-0.5437***		0.3614***
		(0.0572)		(0.0358)
Customer Suggestion		0.2895***		-0.0543
		(0.0538)		(0.0538)
Irrelevant Message		-0.0208		-0.8209***
		(0.0747)		(0.0794)
Number of Observations	10640	10640	10640	10640

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.5: Correlations among key variables (N = 10,640)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.00																		
2	0.28	1.00																	
3	0.02	-0.12	1.00																
4	0.12	0.14	-0.59	1.00															
5	0.01	-0.10	0.84	-0.52	1.00														
6	0.01	0.18	-0.34	0.56	-0.30	1.00													
7	0.01	0.08	-0.15	0.25	-0.14	0.16	1.00												
8	0.21	-0.03	-0.27	0.44	-0.25	-0.21	-0.11	1.00											
9	0.12	0.04	-0.23	-0.21	-0.23	-0.15	-0.07	-0.18	1.00										
10	0.06	-0.02	-0.05	-0.02	-0.09	-0.10	-0.06	0.02	-0.06	1.00									
11	0.03	-0.07	0.03	-0.25	-0.16	-0.17	-0.08	-0.14	-0.14	-0.08	1.00								
12	0.09	0.20	-0.26	0.43	-0.23	0.42	0.24	0.06	-0.05	0.00	-0.21	1.00							
13	0.05	0.05	-0.15	0.20	-0.14	0.14	0.05	0.11	-0.08	0.03	-0.06	0.33	1.00						
14	0.04	0.06	0.02	-0.04	0.01	-0.03	-0.03	-0.02	-0.02	-0.01	0.06	-0.03	0.02	1.00					
15	0.02	0.03	0.02	-0.05	0.00	-0.02	-0.04	-0.04	-0.02	0.03	0.07	-0.09	-0.06	0.00	1.00				
16	0.12	-0.05	-0.08	0.14	-0.08	-0.12	-0.07	0.35	-0.11	0.07	0.00	-0.05	0.02	-0.03	0.60	1.00			
17	0.07	-0.02	0.03	-0.08	0.02	-0.05	-0.02	-0.07	0.04	0.01	0.06	-0.08	-0.04	-0.04	0.28	0.15	1.00		
18	0.00	-0.02	-0.10	0.15	-0.09	0.11	0.13	0.03	-0.05	-0.02	-0.04	0.06	0.03	0.00	0.12	0.05	0.02	1.00	
19	0.10	-0.03	-0.13	0.19	-0.10	0.20	0.18	-0.05	-0.03	-0.02	-0.11	0.10	0.02	-0.02	-0.01	-0.16	0.01	0.62	1.00

Note. 1. Number of Likes; 2. Number of Comments; 3. Positive Valence; 4. Negative Valence; 5. Positive Testimonial; 6. Quality Complaint; 7. Money Complaint; 8. Social Complaint; 9. Customer Question; 10. Customer Suggestion; 11. Irrelevant Message; 12. Log(Word Count); 13. ARI Score; 14. User Activeness; 15. Log(Page Popularity); 16. Log(Post-Level UGC); 17. Log(Post-Level MGC); 18. LexisNexis_1; 19. Log(Assets).

Appendix 2.6: Regression results incorporating 32 user posts with tags of the businesses and 8 user posts that have been shared

	Likes	Likes	Comments	Comments
Log(Word Count)	0.1004***	0.1462***	0.2609***	0.2241***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0101***	0.0078**	-0.0080**	-0.0049
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3262***	-0.3040***	-0.1157	-0.0915
	(0.0513)	(0.0525)	(0.0609)	(0.0599)
Log(Post-Level UGC)	0.2115***	0.1285***	-0.2444***	-0.1931***
	(0.0127)	(0.0132)	(0.0147)	(0.0155)
Log(Post-Level MGC)	-0.1073***	-0.0799***	0.0825***	0.0703***
	(0.0227)	(0.0223)	(0.0167)	(0.0168)
LexisNexis_1	0.0134***	0.0116***	-0.0003	0.0015
	(0.0023)	(0.0023)	(0.0023)	(0.0022)
User Activeness	0.0042***	0.0027**	0.0023*	0.0031**
	(0.0010)	(0.0010)	(0.0011)	(0.0010)
Positive Valence	0.5322***		-0.3483***	
	(0.0488)		(0.0378)	
Negative Valence	0.7687***		0.0540	
	(0.0473)		(0.0337)	
Positive Testimonial		0.2499***		-0.2065***
		(0.0509)		(0.0429)
Quality Complaint		0.1132*		0.2751***
		(0.0470)		(0.0365)
Money Complaint		0.1318*		0.1310**
		(0.0667)		(0.0468)
Social Complaint		0.9262***		-0.2105***
		(0.0502)		(0.0518)
Customer Question		-0.5309***		0.3647***
		(0.0571)		(0.0357)
Customer Suggestion		0.2953***		-0.0519
		(0.0540)		(0.0538)
Irrelevant Message		0.0046		-0.8176***
		(0.0740)		(0.0787)
Number of Observations	10680	10680	10680	10680

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.7: Regression results using an alternative measure of *user activeness*, i.e., the number of posts from a user on a specific business page within the time window of 3 months before the focal post. Posts between January and March are dropped

	Likes	Likes	Comments	Comments
Log(Word Count)	0.1125***	0.1594***	0.2827***	0.2393***
	(0.0168)	(0.0168)	(0.0149)	(0.0159)
ARI Score	0.0096**	0.0070*	-0.0088**	-0.0047
	(0.0033)	(0.0033)	(0.0033)	(0.0032)
Log(Page Popularity)	-0.3348***	-0.3274***	-0.0524	-0.0520
	(0.0590)	(0.0603)	(0.0672)	(0.0666)
Log(Post-Level UGC)	0.1867***	0.1046***	-0.2571***	-0.2015***
	(0.0135)	(0.0140)	(0.0163)	(0.0171)
Log(Post-Level MGC)	-0.1362***	-0.0863*	0.1735***	0.1435***
	(0.0362)	(0.0359)	(0.0341)	(0.0341)
LexisNexis_1	0.0141***	0.0121***	-0.0026	-0.0007
	(0.0024)	(0.0024)	(0.0025)	(0.0025)
User Activeness	0.0106***	0.0072**	0.0045	0.0047
	(0.0024)	(0.0024)	(0.0033)	(0.0032)
Positive Valence	0.5153***		-0.3762***	
	(0.0556)		(0.0435)	
Negative Valence	0.7566***		0.0053	
	(0.0528)		(0.0387)	
Positive Testimonial		0.2144***		-0.1786***
		(0.0562)		(0.0493)
Quality Complaint		0.0735		0.2801***
		(0.0507)		(0.0418)
Money Complaint		0.0625		0.1724**
		(0.0752)		(0.0529)
Social Complaint		0.8967***		-0.2288***
		(0.0539)		(0.0581)
Customer Question		-0.5498***		0.3988***
		(0.0631)		(0.0411)
Customer Suggestion		0.3064***		-0.0338
		(0.0580)		(0.0632)
Irrelevant Message		-0.0526		-0.7969***
		(0.0851)		(0.0936)
Number of Observations	8221	8221	8221	8221

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.8: Regression results using an alternative measure of *page popularity*, i.e., the total number of user- and company-generated posts within the time window of 3 months prior to the focal post. Posts between January and March are dropped

	Likes	Likes	Comments	Comments
Log(Word Count)	0.1124***	0.1605***	0.2834***	0.2400***
	(0.0168)	(0.0169)	(0.0150)	(0.0159)
ARI Score	0.0094**	0.0067*	-0.0090**	-0.0049
	(0.0033)	(0.0033)	(0.0033)	(0.0033)
Log(Page Popularity)	0.0077	-0.0063	-0.0114	0.0062
	(0.0370)	(0.0365)	(0.0408)	(0.0407)
Log(Post-Level UGC)	0.1641***	0.0853***	-0.2590***	-0.2069***
	(0.0140)	(0.0143)	(0.0175)	(0.0182)
Log(Post-Level MGC)	-0.1635***	-0.1130**	0.1716***	0.1412***
	(0.0359)	(0.0355)	(0.0340)	(0.0340)
LexisNexis_1	0.0131***	0.0113***	-0.0028	-0.0009
	(0.0024)	(0.0025)	(0.0025)	(0.0024)
User Activeness	0.0046***	0.0032**	0.0019	0.0023
	(0.0010)	(0.0010)	(0.0012)	(0.0012)
Positive Valence	0.5072***		-0.3772***	
	(0.0555)		(0.0435)	
Negative Valence	0.7682***		0.0065	
	(0.0527)		(0.0387)	
Positive Testimonial		0.1998***		-0.1796***
		(0.0560)		(0.0493)
Quality Complaint		0.0745		0.2806***
		(0.0504)		(0.0418)
Money Complaint		0.0644		0.1722**
		(0.0753)		(0.0530)
Social Complaint		0.9008***		-0.2263***
		(0.0536)		(0.0580)
Customer Question		-0.5516***		0.3988***
		(0.0630)		(0.0411)
Customer Suggestion		0.3105***		-0.0319
		(0.0578)		(0.0631)
Irrelevant Message		-0.0782		-0.8026***
		(0.0845)		(0.0935)
Number of Observations	8221	8221	8221	8221

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.9: Regression results using alternative measures for *post-level UGC* and *post-level MGC*, i.e., the number of user- and company-generated posts only 24 hours before a focal post, respectively

	Likes	Likes	Comments	Comments
Log(Word Count)	0.0987***	0.1462***	0.2655***	0.2269***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0105***	0.0081**	-0.0082**	-0.0048
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3610***	-0.3480***	-0.1685**	-0.1414*
	(0.0522)	(0.0536)	(0.0609)	(0.0600)
Log(Post-Level UGC)	0.2237***	0.1420***	-0.2124***	-0.1605***
	(0.0124)	(0.0129)	(0.0143)	(0.0151)
Log(Post-Level MGC)	-0.0918***	-0.0742**	0.0644***	0.0530**
	(0.0229)	(0.0228)	(0.0179)	(0.0179)
LexisNexis_1	0.0125***	0.0108***	-0.0006	0.0013
	(0.0023)	(0.0023)	(0.0023)	(0.0022)
User Activeness	0.0040***	0.0025**	0.0026*	0.0033**
	(0.0009)	(0.0009)	(0.0011)	(0.0010)
Positive Valence	0.5498***		-0.3547***	
	(0.0490)		(0.0379)	
Negative Valence	0.7844***		0.0341	
	(0.0473)		(0.0337)	
Positive Testimonial		0.2489***		-0.2061***
		(0.0508)		(0.0430)
Quality Complaint		0.1000*		0.2714***
		(0.0469)		(0.0366)
Money Complaint		0.1304		0.1280**
		(0.0666)		(0.0468)
Social Complaint		0.9353***		-0.2500***
		(0.0499)		(0.0520)
Customer Question		-0.5461***		0.3673***
		(0.0571)		(0.0358)
Customer Suggestion		0.2884***		-0.0611
		(0.0538)		(0.0539)
Irrelevant Message		-0.0144		-0.8263***
		(0.0745)		(0.0795)
Number of Observations	10640	10640	10640	10640

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.10: Regression results using two alternative measures, *LexisNexis_7* or *LexisNexis_14*, for *LexisNexis_1*

Use Lexis_7	Likes	Likes	Comments	Comments
Log(Word Count)	0.0997***	0.1467***	0.2610***	0.2245***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0101***	0.0078**	-0.0082**	-0.0050
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3861***	-0.3631***	-0.0948	-0.0768
	(0.0531)	(0.0545)	(0.0615)	(0.0607)
Log(Post-Level UGC)	0.2238***	0.1404***	-0.2423***	-0.1916***
	(0.0126)	(0.0132)	(0.0147)	(0.0155)
Log(Post-Level MGC)	-0.1134***	-0.0862***	0.0811***	0.0693***
	(0.0228)	(0.0224)	(0.0167)	(0.0168)
LexisNexis_7	0.0059***	0.0046***	-0.0024*	-0.0016
	(0.0011)	(0.0011)	(0.0010)	(0.0010)
User Activeness	0.0042***	0.0027**	0.0024*	0.0031**
	(0.0010)	(0.0010)	(0.0011)	(0.0010)
Positive Valence	0.5402***		-0.3535***	
	(0.0490)		(0.0379)	
Negative Valence	0.7735***		0.0528	
	(0.0474)		(0.0337)	
Positive Testimonial		0.2471***		-0.2097***
		(0.0509)		(0.0429)
Quality Complaint		0.1008*		0.2727***
		(0.0469)		(0.0365)
Money Complaint		0.1318*		0.1311**
		(0.0667)		(0.0468)
Social Complaint		0.9280***		-0.2026***
		(0.0502)		(0.0518)
Customer Question		-0.5436***		0.3616***
		(0.0572)		(0.0357)
Customer Suggestion		0.2889***		-0.0547
		(0.0538)		(0.0538)
Irrelevant Message		-0.0146		-0.8232***
		(0.0747)		(0.0793)
Number of Observations	10640	10640	10640	10640

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Use Lexis_14	Likes	Likes	Comments	Comments
Log(Word Count)	0.0995***	0.1468***	0.2608***	0.2244***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0102***	0.0078**	-0.0081**	-0.0049
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3742***	-0.3499***	-0.0750	-0.0571
	(0.0534)	(0.0549)	(0.0616)	(0.0608)
Log(Post-Level UGC)	0.2240***	0.1397***	-0.2425***	-0.1919***
	(0.0127)	(0.0132)	(0.0147)	(0.0155)
Log(Post-Level MGC)	-0.1139***	-0.0867***	0.0809***	0.0692***
	(0.0228)	(0.0225)	(0.0167)	(0.0168)
LexisNexis_14	0.0023***	0.0016*	-0.0021**	-0.0017**
	(0.0007)	(0.0007)	(0.0006)	(0.0006)
User Activeness	0.0043***	0.0027**	0.0024*	0.0031**
	(0.0009)	(0.0010)	(0.0011)	(0.0010)
Positive Valence	0.5411***		-0.3536***	
	(0.0491)		(0.0379)	
Negative Valence	0.7731***		0.0538	
	(0.0474)		(0.0337)	
Positive Testimonial		0.2475***		-0.2095***
		(0.0509)		(0.0429)
Quality Complaint		0.0978*		0.2731***
		(0.0469)		(0.0365)
Money Complaint		0.1298		0.1323**
		(0.0667)		(0.0468)
Social Complaint		0.9345***		-0.1997***
		(0.0502)		(0.0518)
Customer Question		-0.5444***		0.3620***
		(0.0572)		(0.0357)
Customer Suggestion		0.2857***		-0.0544
		(0.0538)		(0.0538)
Irrelevant Message		-0.0132		-0.8233***
		(0.0747)		(0.0793)
Number of Observations	10640	10640	10640	10640

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.11: Regression results using random effects logistic regression model

	Likes	Likes	Comments	Comments
Log(Word Count)	0.1212***	0.1797***	0.4875***	0.4145***
	(0.0225)	(0.0235)	(0.0246)	(0.0259)
ARI Score	0.0156***	0.0108*	-0.0132**	-0.0076
	(0.0047)	(0.0048)	(0.0050)	(0.0052)
Log(Page Popularity)	-0.5027***	-0.4814***	0.3693	0.3805
	(0.1158)	(0.1167)	(0.1986)	(0.2047)
Log(Post-Level UGC)	0.2929***	0.2030***	-0.4221***	-0.3334***
	(0.0228)	(0.0240)	(0.0244)	(0.0258)
Log(Post-Level MGC)	-0.1531***	-0.1212***	0.1536***	0.1274***
	(0.0330)	(0.0332)	(0.0351)	(0.0361)
LexisNexis_1	0.0149***	0.0134***	0.0045	0.0081*
	(0.0037)	(0.0037)	(0.0038)	(0.0040)
User Activeness	0.0040	0.0028	0.0003	0.0030
	(0.0023)	(0.0023)	(0.0023)	(0.0024)
Positive Valence	0.7526***		-0.4611***	
	(0.0653)		(0.0631)	
Negative Valence	1.0088***		-0.0280	
	(0.0632)		(0.0621)	
Positive Testimonial		0.3658***		-0.0920
		(0.0754)		(0.0788)
Quality Complaint		0.1981**		0.7200***
		(0.0717)		(0.0771)
Money Complaint		0.1538		0.1798
		(0.0944)		(0.1083)
Social Complaint		1.2497***		-0.4261***
		(0.0834)		(0.0863)
Customer Question		-0.6377***		1.0638***
		(0.0752)		(0.0791)
Customer Suggestion		0.3380***		0.0958
		(0.0892)		(0.0940)
Irrelevant Message		-0.2122		-1.1236***
		(0.1119)		(0.1186)
Number of Observations	10640	10640	10640	10640

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.12: Regression results after dropping 37 posts with videos

	Likes	Likes	Comments	Comments
Log(Word Count)	0.1003***	0.1466***	0.2625***	0.2255***
	(0.0152)	(0.0153)	(0.0129)	(0.0137)
ARI Score	0.0103***	0.0078**	-0.0085**	-0.0053
	(0.0030)	(0.0030)	(0.0029)	(0.0029)
Log(Page Popularity)	-0.3594***	-0.3425***	-0.1232*	-0.0994
	(0.0525)	(0.0538)	(0.0612)	(0.0603)
Log(Post-Level UGC)	0.2205***	0.1367***	-0.2425***	-0.1917***
	(0.0127)	(0.0133)	(0.0148)	(0.0155)
Log(Post-Level MGC)	-0.1134***	-0.0870***	0.0823***	0.0697***
	(0.0229)	(0.0225)	(0.0168)	(0.0168)
LexisNexis_1	0.0132***	0.0113***	-0.0002	0.0015
	(0.0023)	(0.0023)	(0.0023)	(0.0022)
User Activeness	0.0043***	0.0028**	0.0024*	0.0031**
	(0.0010)	(0.0010)	(0.0011)	(0.0010)
Positive Valence	0.5460***		-0.3538***	
	(0.0492)		(0.0379)	
Negative Valence	0.7782***		0.0499	
	(0.0476)		(0.0337)	
Positive Testimonial		0.2473***		-0.2118***
		(0.0509)		(0.0429)
Quality Complaint		0.1011*		0.2721***
		(0.0470)		(0.0365)
Money Complaint		0.1288		0.1287**
		(0.0667)		(0.0468)
Social Complaint		0.9308***		-0.2110***
		(0.0502)		(0.0519)
Customer Question		-0.5432***		0.3613***
		(0.0572)		(0.0358)
Customer Suggestion		0.2885***		-0.0526
		(0.0539)		(0.0537)
Irrelevant Message		-0.0140		-0.8409***
		(0.0752)		(0.0801)
Number of Observations	10603	10603	10603	10603

Note. * p<0.05; ** p<0.01; *** p<0.001. Standard errors in parentheses. Omitted estimates for constant, industry dummies, type dummies, and asset for brevity.

Appendix 2.13: Exploratory online survey

Question	Response Type	Response Options	Source
We would like to ask about your experience with a specific company's Facebook page. Think of a company whose Facebook page you are most familiar with or have visited most frequently. Copy and paste the URL of the company's Facebook page below.	Text input	NA	NA
What is your relationship with this company?	Multiple choice	• I have purchased products or services from this company • I am interested in purchasing products or services from this company • I work for this company • Other. Please specify	NA
How frequently do you visit this company's Facebook page?	Multiple choice	• Daily • Weekly • Monthly • Quarterly • Yearly • Very Rarely	NA
Why do you VISIT this company's Facebook page? The following statements describe a list of possible reasons. Please indicate the degree to which you agree with each statement. 1. Because I enjoy learning about the company's products or services 2. Because it is fun to check out what happens with the company 3. Because browsing the company's Facebook page is pleasant 4. Because I enjoy learning about other users' experience with the company	5-point Likert scale	• Strongly disagree • Somewhat disagree • Neither agree nor disagree • Somewhat agree • Strongly agree	Adapted from McAlexander et al. (2002)

5. Because it is fun to interact and exchange information about this company with other Facebook users 6. Because chatting with other Facebook users on the page is pleasant 7. Because visiting the company's page is useful to me 8. Because the company's page provides me with useful information 9. Because I visit the page to receive financial benefits from the company 10. Because I visit the page to communicate with the company about a particular issue 11. Because learning about other users' experience with the company is useful to me 12. Because other users on the page provide me with useful information 13. Because I visit the page to ask other users to help me with a particular issue 14. Because I want to socialize with employees of the company 15. Because I want to interact with the social media staff of the company 16. Because I'm a loyal customer of the company 17. Because I feel emotionally connected with the company 18. Because I want to socialize with other users on the page 19. Because I want to interact with friends of mine on the page 20. Because I meet nice people on the page 21. Because I want to support the company 22. Because I like helping the company on its Facebook page 23. Because I want to help the company to be successful 24. Because I want to support the user community 25. Because I like helping other users on the company's Facebook page			

26. Because I want to help other users to solve their problems			
How often do you engage in each of the following activities? 1. Reading posts written by other Facebook users on the page 2. Reading posts written by the company on the page 3. Posting on the page 4. Liking posts written by other Facebook users on the page 5. Liking posts written by the company on the page 6. Commenting on posts written by other Facebook users on the page 7. Commenting on posts written by the company on the page	Multiple choice	• Daily • Weekly • Monthly • Quarterly • Rarely • Never	NA
Why do you READ posts written by other Facebook users on the page? The following statements describe a list of possible reasons. Please indicate the degree to which you agree with each statement. 1. Because posts written by other users help me make the right decisions related to the company 2. Because I benefit from learning others' experiences before I buy a good or service 3. Because other users' posts give me fast information about the company 4. Because other users' posts give me credible information about the company 5. Because I can see if I am the only one who thinks of the company in a certain way 6. Because I like to compare my evaluation of the company with other users' 7. Because I feel much better when I read that I am not the only one who has a certain problem 8. Because I like being part of the user community	5-point Likert scale	• Strongly disagree • Somewhat disagree • Neither agree nor disagree • Somewhat agree • Strongly agree	Adapted from Hennig-Thurau and Walsh (2003)

9. Because I enjoy participating in this user community 10. Because I want to learn about what's happening in the user community 11. Because other users' posts provide me the right answers when I have questions or difficulties with a product or service 12. Because other users' posts provide me advice and solutions to my problems 13. Because the posts are written by other users that I frequent interact with 14. Because the posts are written by other customers of the company 15. Because the posts are written by members of the user community 16. Because other users' posts are fun to read 17. Because I enjoy reading other users' posts 18. Because reading other users' posts helps me kill time			
Why do you POST on the company's page? The following statements describe a list of possible reasons. Please indicate the degree to which you agree with each statement. 1. Because when I publicize the matter on Facebook, companies are more accommodating 2. Because It is more convenient to post on Facebook than writing to or calling the company 3. Because one has more power together with others on Facebook than writing a single letter of complaint 4. Because I want to get anger off my chest 5. Because I want to take vengeance upon the company 6. Because the company harmed me, and now I will harm the company 7. Because my posts help me shake off frustration about bad experiences 8. Because I want to help others by sharing my positive experiences	5-point Likert scale	• Strongly disagree • Somewhat disagree • Neither agree nor disagree • Somewhat agree • Strongly agree	Adapted from Hennig-Thurau et al. (2004); Nimako et al. (2012)

9. Because I want to help other users buy the right products or services 10. Because I want to warn other users about bad products or services 11. Because I want to save others from having the same negative experiences 12. Because I want to raise important corporate social responsibility issues among Facebook users 13. Because I can express my joy about a good experience 14. Because I can tell other users about a great experience 15. Because I feel good when I can tell others my buying success 16. Because my posts show others that I am a clever customer 17. Because a chat with like-minded people is a nice thing for me 18. Because it is fun to communicate with other users on the page 19. Because it is fun to communicate with employees of the company on the page 20. Because I want to share my feedback to a particular employee in the company 21. Because I meet and interact with nice people on the page 22. Because I receive incentives like coupons or discounts 23. Because I get rewards for posting 24. Because I post to support a good company 25. Because I am satisfied with the company and want to help it succeed 26. Because I want to make a suggestion to help the company with its products, services, social responsibility issues, etc. 27. Because I hope to receive advice from others to help solve my problems 28. Because I want to get tips or support from the company			

29. Because I want to ask a question about the company' products, services, or other issues 30. Because I want to receive tips or support from other users 31. Because I want to seek corrective actions from the company about a bad experience 32. Because I want to seek explanations or apologies from the company about a bad experience 33. Because I want to seek remedy or compensation from the company about a bad experience			
Why do you LIKE posts written by other Facebook users on the page? The following statements describe a list of possible reasons. Please indicate the degree to which you agree with each statement. 1. Because I agree with what the users were saying 2. Because I agree with the content of the posts 3. Because I have had experiences similar to those of the users who posted the messages 4. Because I share the feelings of the users who posted the messages 5. Because I want to express my support to the users who posted the messages 6. Because I want other users to know that I support what they were saying 7. Because I want other users to know that I pay attention to their posts 8. Because I want to show that I care about what other users were saying 9. Because the users had liked my posts before 10. Because I want to return the favor of other users who had liked my posts before 11. Because I find the content of the posts interesting 12. Because liking others' posts on the page is a fun thing to do 13. Because I personally know the users who posted the messages	5-point Likert scale	• Strongly disagree • Somewhat disagree • Neither agree nor disagree • Somewhat agree • Strongly agree	Adapted from Scissors et al. (2016)

14. Because the users who posted the messages were my friends 15. Because liking is a nice way of interacting with other users on the page 16. Because I want to acknowledge other users' contribution to this community			
Why do you COMMENT on posts written by other Facebook users on the page? The following statements describe a list of possible reasons. Please indicate the degree to which you agree with each statement. 1. Because I agree with the posts 2. Because I want to express my support to the posts 3. Because I want to express my positive opinions and thoughts about what the users were saying 4. Because I disagree with the posts 5. Because I want to argue against the posts 6. Because I want to share my negative opinions and thoughts about what the users were saying 7. Because I want to add to the discussion by sharing my experience 8. Because I want to ask for clarifications 9. Because I want to follow up on what the users were saying 10. Because I want to raise awareness of the issues mentioned in the posts 11. Because I find the content of the posts interesting 12. Because commenting on others' posts is a fun thing to do on the page 13. Because commenting makes me feel less lonely 14. Because by commenting, I won't have to feel alone 15. Because commenting is a nice way of interacting with other users on the page 16. Because I want to interact with other users on the page	5-point Likert scale	• Strongly disagree • Somewhat disagree • Neither agree nor disagree • Somewhat agree • Strongly agree	Adapted from Smock et al. (2011)

17. Because I personally know the users who posted the messages 18. Because the users who posted the messages were my friends 19. Because I want to answer other users' questions 20. Because I want to help other users with their problems by replying to their posts			
What is your gender?	Multiple choice	• Female • Male	NA
What is your age?	Multiple choice	• Under 25 years old • 25 - 34 years old • 35 - 44 years old • 45 - 54 years old • 55 years old and over • I prefer not to say	NA

Appendix 2.14: Top motivations reported by survey participants for each type of behavior

For each survey item, we calculated the average reported score (5-point Likert scale, 1 – strongly disagree, 5 – strongly agree) across participants who have visited the business pages of Fortune-500 companies (i.e., our research context). For each type of behavior, we list the 5 items that received the highest average scores, indicating the top 5 most prevalent motivations.

Behavior	Top 5 Reported Motivations	Average Score
Visit business pages	• I enjoy learning about the company's products or services;	4.27
	• It is fun to check out what happens with the company;	3.59
	• Browsing the company's Facebook page is pleasant;	3.68
	• I enjoy learning about other users' experience with the company;	3.45
	• It is fun to interact and exchange information with other Facebook users.	3.22
Read user posts	• To benefit from others' experiences before I buy a good or use a service;	3.86
	• Because I like to compare my own evaluation of the company with that of others;	3.59
	• Because I really like being part of such a community of users;	3.45
	• Because I enjoy participating in this user community;	3.45
	• To find advice and solutions for my problems.	3.45
Post on business pages	• This way I can express my joy about a good experience;	3.95
	• I can tell others about a great experience;	3.79
	• I want to ask a question about the company' products, services, or other issues;	3.79
	• I want to give others the opportunity to buy the right products;	3.58
	• I want to make a suggestion to help the company about its products, services, social responsibility issues, etc.	3.53
Like user posts	• I agree with the content of the posts;	4.05
	• I agree with what the users were saying;	3.90
	• I find the content of the posts interesting;	3.80
	• I have had similar experiences as the users who posted the messages;	3.65
	• I share the feelings with the users who posted the messages.	3.50
Comment on user posts	• I agree with the content of the posts;	3.82
	• I find the content of the posts interesting;	3.76
	• I want to add to the discussion by sharing my experience;	3.65
	• I want to share my positive opinions and thoughts about what the users were saying;	3.47
	• I want to answer other users' questions.	3.47

Appendix 4.1: Proof for estimation bias in linear regression with a single misclassified regressor (for reference, see Gustafson 2003)

Suppose the regression equation is $Y = \beta_0 + \beta_1 X + \varepsilon$. Instead of true value X we observe $\hat{X}$, which has misclassification. According to law of iterative expectation, $E(Y|\hat{X}) = E(E(Y|X,\hat{X})|\hat{X})$. Additionally, the nondifferential misclassification assumption implies that $E(Y|X,\hat{X}) = E(Y|X)$. Combining them together, we have the following relationship:

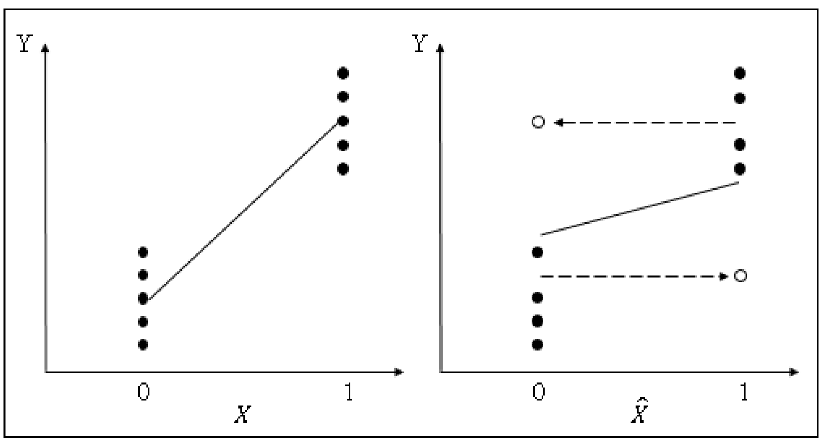
$$E(Y|\hat{X}) = E(E(Y|X)|\hat{X}) = E(\beta_0 + \beta_1 X|\hat{X}) = \beta_0 + \beta_1 E(X|\hat{X})$$

Therefore, $E(\widehat{\beta_1}|\hat{X}) = E(Y|\hat{X}=1) - E(Y|\hat{X}=0) = \beta_1[E(X|\hat{X}=1) - E(X|\hat{X}=0)]$

Further, $E(X|\hat{X}=1) = 1 \times \Pr(X=1|\hat{X}=1) + 0 \times \Pr(X=0|\hat{X}=1) = \Pr(X=1|\hat{X}=1)$. Similarly, $E(X|\hat{X}=0) = \Pr(X=1|\hat{X}=0)$. As a result, we have:

$$E(\widehat{\beta_1}|\hat{X}) = \beta_1[\Pr(X=1|\hat{X}=1) - \Pr(X=1|\hat{X}=0)]$$

Appendix 4.2: A Graphical Illustration of Estimation Bias due to Misclassification



Note. Consider a linear regression of *Y* on a dummy variable *X*. This graph shows the fitted regression line with 10 data points. In the subgraph on the left, all data is correctly measured. In the subgraph on the right, one data point in each class is misclassified as having the opposite class label (corresponding to 80% precision for both class 0 and class 1). Change in the slope of the regression line demonstrates the bias due to misclassification in independent variable. In this case, misclassification in *X* would result in a coefficient that is only 60% of its true value in expectation.

Appendix 4.3: Pseudocode for implementing SIMEX and MC-SIMEX methods (for reference, see Cook and Stefanski 1994 and Küchenhoff et al. 2006).

Given a data set $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$ and the regression model $\boldsymbol{Y} = \boldsymbol{\beta}[\boldsymbol{X}\ \boldsymbol{Z}] + \boldsymbol{\varepsilon}$, we consider $\boldsymbol{X}$ to be the variable that has measurement error or misclassification, and $\boldsymbol{Z}$ to be other precisely measured variables. Here is the pseudocode for estimating error-corrected $\boldsymbol{\beta_{simex}}$ and $\boldsymbol{\beta_{mcsimex}}$, respectively.

Algorithm: Pseudocode for Implementing SIMEX
$\boldsymbol{X}$ has measurement error with standard deviation σ_e, i.e., $X = X_{true} + e$ and $Var(e) = \sigma_e^2$

// Simulation Step:
For λ from $\{\lambda_1, \lambda_2, \dots \lambda_m\}$: // Construct simulated data, $\{\lambda_1, \lambda_2, \dots \lambda_m\}$ can be $\{1,2,\dots,m\}$
For iteration *i* from 1 to *B*:
Generate $\boldsymbol{X}(\lambda)_i$ as $\boldsymbol{X}(\lambda)_i = \boldsymbol{X} + \sqrt{\lambda_i}\sigma_e \boldsymbol{z}$, $\boldsymbol{z} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$
Assemble a new data set $(\boldsymbol{Y}, \boldsymbol{X}(\lambda)_i, \boldsymbol{Z})$
Estimate $\boldsymbol{\beta}(\lambda)_i$ from regression model
Calculate $\boldsymbol{\beta}(\lambda) = B^{-1} \sum_{i=1}^{B} \boldsymbol{\beta}(\lambda)_i$
 	
// Extrapolation Step:
Fit a parametric model over $\{\boldsymbol{\beta}(\lambda_1), \boldsymbol{\beta}(\lambda_2), \dots, \boldsymbol{\beta}(\lambda_m)\}$
Extrapolate to $\boldsymbol{\beta}(-1)$
Obtain $\boldsymbol{\beta_{simex}} = \boldsymbol{\beta}(-1)$

Algorithm: Pseudocode for Implementing MC-SIMEX
$\boldsymbol{X}$ has misclassification, described by the misclassification matrix $\boldsymbol{\Pi}$.

// Simulation Step:
For λ from $\{\lambda_1, \lambda_2, \dots \lambda_m\}$: // Construct simulated data, $\{\lambda_1, \lambda_2, \dots \lambda_m\}$ can be $\{1,2,\dots,m\}$
For iteration *i* from 1 to *B*:
Generate $\boldsymbol{X}(\lambda)_i$ with misclassification of magnitude $\boldsymbol{\Pi}^{(1+\lambda)}$
Assemble a new data set $(\boldsymbol{Y}, \boldsymbol{X}(\lambda)_i, \boldsymbol{Z})$
Estimate $\boldsymbol{\beta}(\lambda)_i$ from regression model
Calculate $\boldsymbol{\beta}(\lambda) = B^{-1} \sum_{i=1}^{B} \boldsymbol{\beta}(\lambda)_i$
 	
// Extrapolation Step:
Fit a parametric model over $\{\boldsymbol{\beta}(\lambda_1), \boldsymbol{\beta}(\lambda_2), \dots, \boldsymbol{\beta}(\lambda_m)\}$
Extrapolate to $\boldsymbol{\beta}(-1)$
Obtain $\boldsymbol{\beta_{mcsimex}} = \boldsymbol{\beta}(-1)$

Appendix 4.4: Diagnostic regression analysis for real-world example in Section 4.5.1 (N = 2,391)

	LP Model			Logit Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1630***	0.1477***	0.1746***	-1.6315***	-1.7172***	-1.5737***
	(0.0206)	(0.0216)	(0.0294)	(0.1324)	(0.1407)	(0.1811)
Sentiment	**-0.0855*** **	**-0.0628** **	**-0.0890** **	**-0.5300*** **	**-0.3912** **	**-0.5304** **
	(0.0192)	**(0.0193)**	**(0.0288)**	**(0.1227)**	**(0.1237)**	**(0.1774)**
Photo	-0.02545	-0.0251	-0.0243	-0.1927	-0.1909	-0.1832
	(0.0138)	(0.0139)	(0.0139)	(0.1209)	(0.1223)	(0.1211)
Words	0.9275***	0.8890***	0.8185***	5.1059***	4.8561***	4.4318***
	(0.0958)	(0.0989)	(0.1044)	(0.6005)	(0.6199)	(0.6525)
Sequence	0.0441	0.0414	0.0380	0.2945	0.2783	0.2491
	(0.0329)	(0.0329)	(0.0330)	(0.2143)	(0.2147)	(0.2158)
Log Likelihood	-1069.4	-1073.9		-1091.6	-1095.7	
AIC	2148.7	2157.8		2193.2	2201.4	

Note. This table contains regression results using the labeled dataset, i.e., 20% (or 2,391) of all reviews.

Appendix 4.5: R code used for MC-SIMEX correction in Section 4.5.1

```
library(simex)        # Attach the “simex” library.

data = read.csv()     # Read in the dataset that contains all variables and sentiment prediction for
                        each review.

mc = matrix(c(0.74,0.26,0.07,0.93), nrow = 2)  # Specify the misclassification matrix.
dimnames(mc) = list(c("0", "1"), c("0", "1"))    # Assign the class label as dimension names of
                                                   the misclassification matrix.

# Running linear regressions.
# Specify the “family” parameter in glm() to run other types of regressions.
# First, run a linear regression with true values of sentiment and control variables.
# Note that this step does not exist in actual studies, because true values are not observed.

model.t = glm(helpfulness ~ true_sentiment + control_variables, data = data)
summary(model.t)

# Second, run a linear regression with predicted values of sentiment and control variables.

model.mc = glm(helpfulness ~ predicted_sentiment + control_variables, data = data)
summary(model.mc)

# Third, specify the regression that contains misclassification. Specify parameters “x = T, y = T”
to inform glm() to # return the response vector and model matrix used in model fitting.

naive = glm(helpfulness ~ predicted_sentiment + control_variables, data = data, x = T, y = T)

# Finally, perform MC-SIMEX correction by calling the mcsimex() function. The first input is
the regression with misclassification. The second parameter “SIMEXvariable” specifies the
name of the variable with error. The third parameter “mc.matrix” specifies the misclassification
matrix. For other parameters, please see the manual for mcsimex() function.

model.simex = mcsimex(naive, SIMEXvariable = " predicted_sentiment ", mc.matrix = mc)
summary(model.simex)
```

Appendix 4.6: Additional analyses for real-world example in Section 4.5.1 with different sample sizes.

Regression Results and Corrections of the TripAdvisor.com Dataset (**N = 500**)

	LP Model			Logit Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1482***	0.1074*	0.1158	-1.7203***	-1.9786***	-1.8926***
	(0.0430)	(0.0482)	(0.0702)	(0.2862)	(0.3364)	(0.4592)
Sentiment	**-0.0692**	**-0.0162**	**-0.0244**	**-0.4655**	**-0.1079**	**-0.1978**
	(0.0408)	**(0.0431)**	**(0.0702)**	**(0.2731)**	**(0.2964)**	**(0.4594)**
Photo	0.0441	0.0413	0.0412	0.2555	0.2344	0.2360
	(0.0319)	(0.0319)	(0.0395)	(0.1852)	(0.1845)	(0.1786)
Words	0.9571***	0.9742***	0.9510**	5.5621***	5.6385***	5.4612***
	(0.2257)	(0.2342)	(0.2900)	(1.4218)	(1.4910)	(1.6467)
Sequence	-0.0394	-0.0503	-0.0493	-0.3060	-0.3873	-0.3756
	(0.0761)	(0.0760)	(0.0838)	(0.5785)	(0.5769)	(0.6726)
Log Likelihood	-212.49	-213.86		-220.73	-222.06	
AIC	436.97	439.72		451.45	454.12	

Note. The MC-SIMEX method does not provide log likelihood or AIC statistics.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.1$

Regression Results and Corrections of the TripAdvisor.com Dataset (**N = 2,000**)

	LP Model			Logit Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1763***	0.1452***	0.1660***	-1.5464***	-1.7215***	-1.6175***
	(0.0232)	(0.0245)	(0.0353)	(0.1447)	(0.1576)	(0.2109)
Sentiment	**-0.0803*** **	**-0.0411+**	**-0.0618+**	**-0.4819*** **	**-0.2475+**	**-0.3452+**
	(0.0215)	**(0.0219)**	**(0.0349)**	**(0.1330)**	**(0.1383)**	**(0.2065)**
Photo	0.0106	0.0074	0.0088	0.0586	0.0382	0.0475
	(0.0195)	(0.0195)	(0.0237)	(0.1181)	(0.1177)	(0.1332)
Words	0.7572***	0.7628***	0.7112***	4.1086***	4.1178***	3.7970***
	(0.1046)	(0.1081)	(0.1439)	(0.6328)	(0.6550)	(0.8094)
Sequence	0.0360	0.0329	0.0334	0.2475	0.2283	0.2299
	(0.0369)	(0.0370)	(0.0391)	(0.2379)	(0.2379)	(0.2420)
Log Likelihood	-933.05	-938.22		-942.67	-947.42	
AIC	1878.10	1888.40		1895.30	1904.80	

Note. The MC-SIMEX method does not provide log likelihood or AIC statistics.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.1$

Regression Results and Corrections of the TripAdvisor.com Dataset (**N = 5,000**)

	LP Model			Logit Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1648***	0.1507***	0.1739***	-1.6065***	-1.6878***	-1.5641***
	(0.0146)	(0.0154)	(0.0229)	(0.0931)	(0.0998)	(0.1346)
Sentiment	**-0.0611*** **	**-0.0416** **	**-0.0651** **	**-0.3740*** **	**-0.2544** **	**-0.3770** **
	(0.0135)	**(0.0137)**	**(0.0222)**	**(0.0848)**	**(0.0869)**	**(0.1293)**
Photo	-0.0255*	-0.0258*	-0.0250**	-0.1986*	-0.2008*	-0.1928*
	(0.0107)	(0.0107)	(0.0095)	(0.0902)	(0.0902)	(0.0924)
Words	0.8948***	0.8834***	0.8267***	4.9458***	4.8607***	4.5130***
	(0.0703)	(0.0725)	(0.0938)	(0.4312)	(0.4457)	(0.5312)
Sequence	-0.0353	-0.0385+	-0.0379	-0.2197	-0.2402	-0.2405
	(0.0228)	(0.0228)	(0.0235)	(0.1581)	(0.1582)	(0.1682)
Log Likelihood	-2315.62	-2321.24		-2343.63	-2348.86	
AIC	4643.20	4654.50		4697.30	4707.70	

Note. The MC-SIMEX method does not provide log likelihood or AIC statistics.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.1$

Appendix 4.7: Diagnostic regression analysis for real-world example in Section 4.5.2 (N = 806)

	OLS Model			Poisson Model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	-0.6257	-1.0333	-0.8095	-1.6350***	-1.7096***	-1.6102***
	(1.0350)	(1.0275)	(1.0407)	(0.3082)	(0.3092)	(0.3393)
Log(Words)	0.3575***	.3945***	0.3308**	0.4494***	0.4628***	0.4310***
	(0.0886)	(0.0909)	(0.1018)	(0.0253)	(0.0266)	(0.0431)
Activeness	0.0676***	0.0662***	0.0654***	0.0150***	0.0144***	0.0150***
	(0.0122)	(0.0122)	(0.0122)	(0.0033)	(0.0033)	(0.0034)
Log(Popularity)	0.1167	0.1369	0.1434	0.0668*	0.0647*	0.0711*
	(0.1003)	(0.1007)	(0.1008)	(0.0307)	(0.0308)	(0.0313)
Type = Link	-0.9259	-1.0160	-1.0751	-1.2153*	-1.1987*	-1.2049*
	(0.6966)	(0.6999)	(0.7000)	(0.5009)	(0.5009)	(0.5019)
Type = Photo	0.0768	-0.1046	0.0371	-0.6038*	-0.7347**	-0.6800*
	(0.5452)	(0.5435)	(0.5519)	(0.2769)	(0.2755)	(0.2875)
Sentiment	**-0.6690****	**-0.4113**	**-0.6430***	**-0.3272*****	**-0.1578***	**-0.3514**
	(0.2114)	**(0.2217)**	**(0.2982)**	**(0.0713)**	**(0.0803)**	**(0.2873)**
Log Likelihood	-1905.6	-1908.9		-2290.4	-2299.5	
AIC	3825.2	3831.8		4594.7	4613.0	

Note. This table contains regression results using the labeled dataset, i.e., 30% (or 410) of all profile pictures. In our diagnostic analysis, the dummy variable *Type = Video* was not estimated, because no video-typed post was selected into the 30% random sample.