Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook*

Dokyun Lee — The Wharton School
Kartik Hosanagar — The Wharton School
Harikesh S. Nair — Stanford GSB

First version: September 2013. This version: September 2015.

Abstract

We investigate the effect of social media advertising content on customer engagement using a large-scale field study on Facebook. We content-code more than 100,000 unique messages across 800 companies using a combination of Amazon Mechanical Turk and state-of-the-art Natural Language Processing algorithms. We use this large-scale dataset of content attributes to test the effect of social media marketing content on subsequent user engagement − defined as *Likes*, comments, shares, and click-throughs − with the messages. We develop methods to account for potential selection biases that arise from Facebook's filtering algorithm, EdgeRank, that assigns messages non-randomly to users. We find that inclusion of widely used content related to brand-personality − like humor, emotion and discussion of the brand's philanthropic positioning − *increases* consumer engagement with a message. We find that directly informative content − like mentions of prices, availability, and product features − *reduce* engagement when included in messages in isolation, but *increase* engagement when provided in combination with brand-personality related attributes. We also find certain directly informative content such as the mention of deals and promotions drive consumers' path-to-conversion (click-throughs). Our results suggest therefore that there may be substantial gains from content engineering by combining informative characteristics associated with immediate leads (via improved click-throughs) with brand personality related content that help maintain future reach and branding on the social media site (via improved engagement). Our results inform content design strategies in social media, and the methodology we develop to content-code large-scale textual data provides a framework for future studies on unstructured natural language data such as advertising content or product reviews.

Keywords: consumer engagement, social media, advertising content, marketing communication, large-scale data, natural language processing, machine learning, selection, Facebook, EdgeRank.

*We thank seminar participants at the ISS Conference (Jan 2013), Mack Institute Conference (Spring 2013), SCECR Conference (Summer 2013), WITS Conference (Dec 2013), INFORMS Conference (Oct 2014), and WISE conference (Dec 2014) for comments, and a collaborating company that wishes to be anonymous for providing the data used in the analysis. The authors gratefully acknowledge the financial support from the Jay H. Baker Retailing Center and Mack Institute of the Wharton School and the Wharton Risk Center (Russell Ackoff Fellowship). The authors also thank David Bell, Jonah Berger, Cexun Jeffrey Cai, Pradeep Chintagunta, Pedro Gardete, Arun Gopalakrishnan, Raghuram Iyengar, Carl Mela, Navdeep Sahni, Olivier Toubia and Christophe Van den Bulte for their helpful feedback. All errors are our own.

1 Introduction

Social networks are increasingly taking up a greater share of consumers' time spent online. As a result, social media − which includes advertising on social networks and/or marketing communication with social characteristics − is becoming a larger component of firms' marketing budgets. As firms increase their social media activity, the role of *content engineering* has become increasingly important. Content engineering seeks to develop content that better engages targeted users and drives the desired goals of the marketer from the campaigns they implement. This raises the question: what content works best? Surprisingly, the answer is not obvious. The most important body of academic work on this topic is the applied psychology and consumer behavior literature which has discussed ways in which the content of marketing communication engages consumers and captures attention. However, most of this work has tested and refined theories about content primarily in laboratory settings. Relatively little has been explored systematically about the empirical consequences of advertising and promotional content in *real-world, field settings* outside the laboratory. Despite its obvious relevance to practice, marketing and advertising content is also relatively under emphasized in economic theory. Canonical models of advertising in which ad-intensity acts as a signal of quality (c.f. Nelson (1974); Kihlstrom and Riordan (1984); Milgrom and Roberts (1986)) do not postulate a role for ad content because advertising intensity conveys all relevant information about product quality in equilibrium to market participants. Economic models of directly informative advertising (c.f. Butters (1977); Grossman and Shapiro (1984)) typically allow for advertising to inform agents only about price and product existence − yet, casual observation and several studies in lab settings (c.f. Armstrong (2010); Berger (2012)) suggest that advertisements contain much more information and content beyond prices. In this paper, we explore the role of content in driving consumer engagement in social media in a large-scale field setting. We document the kinds of content used by firms in practice. We show that a variety of emotional, philanthropic, and directly informative advertising content attributes affect engagement and that the role of content varies significantly across firms and industries. The richness of our engagement data and the ability to content code social media messages in a cost-efficient manner enables us to study the problem at a larger scale than much of the previous literature on the topic.

Our analysis is of direct relevance to industry in better understanding and improving firms' social media marketing strategies. Many industry surveys (Ascend2, 2013; Gerber, 2014; eMarketer, 2013; SmartBrief, 2010; Ragan and Solutions, 2012) report that achieving engagement on large audience platforms like Facebook is an important marketing goal for consumer-facing firms. Social media marketing agencies' financial arrangements are also increasingly contracted on the basis of the engagement these agencies promise to drive for their clients. In the early days of the industry, it was thought that engagement was primarily driven by the volume of users socially connected to the brand by increasing the *reach* of posts released by the firms. Accordingly, firms aggressively acquired fans and followers on platforms like Facebook by investing heavily in ads on the network. However, early audits of the data (e.g., Creamer 2012) suggested that only about 1% of an average firm's Facebook fans show any engagement with the brand by *Liking,* sharing, or commenting on messages by the brand on the platform. This implies poor virality or social reach associated with posts by firms. The development of filtering algorithms by platforms like Facebook, in which the current engagement of users with a firms' posts determine their future reach on the platform, further accentuated the importance of creating engaging content. As a result, industry attention shifted from acquisition of social media followers *per se*, to the design of content that achieves both better reach and engagement amongst social media followers. In a widely reported example that reflected this trend, during the time-span of our data (2011-12), General Motors curtailed its annual spending of $10M on Facebook's paid ads −a vehicle for acquiring new fans for the brand − choosing instead to focus on creating content for its branded Facebook Page, on which it spent $30M (WSJ, 2012). While attention in industry has shifted towards content in this manner, industry still struggles with understanding what kinds of content work better for which firms and in what ways. For example, are messages seeking to inform consumers about product or price attributes more effective than persuasive messages with humor or emotion? Do messages explicitly soliciting user response (e.g., "*Like* this post if ...") draw more engagement or in fact turn users away? Does the same strategy apply across different industries? Our paper systematically explores these kinds of questions and contributes to the formulation of better content engineering policies in practice.[1]

Our empirical investigation is implemented on Facebook, which is the largest social media platform in the world. Many top brands maintain a Facebook page from which they serve posts and messages to connected users. This is a form of free social media marketing that has increasingly become a popular and important channel for marketing. Our data comprises information on about 100,000 such messages posted by a panel of about 800 firms over a 11-month period between September 2011 and July 2012. For each message, our data also contains time-series information on two kinds of engagement measures − *Likes* and comments − observed on Facebook. In addition, we have cross-sectional data on shares and click-throughs. We supplement these engagement data with message attribute information that we create using a large-scale survey we implement on Amazon Mechanical Turk (henceforth "AMT"), combined with a Natural Language Processing algorithm (henceforth "NLP"), we build to tag messages. We incorporate new methods and procedures to improve the accuracy of content tagging which we believe will be useful in future studies analyzing other kinds of advertising content and product reviews.

Our data has several advantages that facilitate a detailed study of content. First, Facebook messages have rich content attributes (unlike say, Twitter tweets, which are restricted to 140 characters) and rich data on user engagement. Second, Facebook requires real names and, therefore, data on user activity on Facebook is often more reliable compared to other social media sites. Third, engagement is measured on a daily basis (panel data) by actual message-level engagement such as *Likes* and comments that are precisely tracked within a closed system. These aspects make Facebook an almost ideal setting to study the role of

[1]As of December 2013 (when this paper was written), industry-leading social media analytics firms such as Wildfire (now part of Google) did not offer detailed content engineering analytics connecting a wide variety of social media content with real engagement data. Rather, they provided simpler analytics such as reporting engagement by time-of-the-day or day-of-the-week to post and split by inclusion of pictures or videos. More recently, the content engineering industry has mushroomed and become more sophisticated in its use of analytics. Content engineering to obtain better reach on Facebook via firm's pages has parallels to Search Engine Optimization (SEO) for obtaining improved organic listings on search engines.

content in social media marketing.

Our strategy for coding content is motivated by the psychology, marketing and economic literatures on advertising (see Cialdini (2001); Bagwell (2007); Berger (2012); Chandy et al. (2001); Vakratsas and Ambler (1999) for some representative overviews). In the economics literature, it is common to classify advertising as informative (shifting beliefs about product existence or prices) or persuasive (shifting preferences directly). This classification is difficult to operationalize in many contexts because of two reasons. Firstly, the basis of informative content is typically limited to prices and/or existence, and exactly what is "persuasive content" is usually not well-defined and is treated as a "catch-all" without finer classification. Secondly, some content can be both "persuasive" and indirectly "informative" (e.g., Sahni et al. (2015)). For instance, the fact that many people in a consumer's social group are using the product may persuade the consumer to use the product (i.e., it is "persuasive") or provide a signal that it's a good match with consumers like him (i.e., it is "*indirectly* informative").

To avoid this coarse distinction and the associated difficulties in interpretation, we first present the full results for the effect each content attribute separately so the reader can judge the disaggregate effects without the need for any ex-post grouping. Secondly, we present a grouping of the content attributes that reflect the type of content we see on Facebook and parallels finer classification schema that has appeared in the advertising literature. Here, we follow the important early work of Resnik and Stern (1977), who operationalize directly informative advertising based on the number of informational cues present in a message (see Abernethy and Franke, 1996 for an overview of studies in this stream). Some criteria that Resnik and Stern (1977) suggest for classifying content as directly informative are whether it includes details about products, promotions, availability, price, and product related aspects that could be used in optimizing the purchase decision. Following this stream, any product oriented facts, and brand and product mentions are categorized as directly informative content.

The other types of content we see in Facebook posts involve aspects of brand personality. For instance, we see thousands of posts from firms that contain humor, emotional appeal, casual banter or discuss the brand's philanthropic outreach. We interpret these as attempts by the firm to establish a brand personality − i.e., "a set of human characteristics associated with the brand." (Aaker 1997; Weiss and Huber 2000). One reason firms may be using such content is because consumers tend to choose brands that are in congruence with their own personalities (Govers and Schoormans 2005). Further, the branding literature suggests that functional benefits of a brand also become more persuasive when expressed by the brand's personality (Keller 1993; Aaker 1996). Overall, we see the role of this type of content as attempts by firms to promote relationship-building and to persuade consumers to use their brand via such relationships.

Estimation of the effect of content on subsequent engagement is complicated by the non-random allocation of messages to users. A typical concern in empirical work in such settings is of reverse causality − that firms target specific content to selected sub-audiences − so, the subsequent covariation in outcomes reflect both the effect of the content as well as the targeting policies of the firms (e.g., Nair et al. 2013). This concern is not first-order in our context because unlike Facebook's banner advertisements or sponsored posts, the Facebook organic page environment does not allow companies to target specific audiences. That is, a firm's posts are meant for all its fans. Instead, all targeting is implemented *ex post* by Facebook via its proprietary EdgeRank algorithm, the goal of which is to present the user with a positive experience on Facebook that is not polluted with content that he does not value. EdgeRank tends to serve to users messages that are newer and are expected to appeal better to his/her tastes. Hence, the main concern for inference created by targeting arises from the selection induced by EdgeRank. We account for the selection induced by EdgeRank by developing a semi-parametric correction for the filtering it induces, and show how it can be incorporated into the estimation procedure. Our correction serves as a semi-parametric "control function" to the non-random selection induced by the filtering algorithm. We discuss other endogeneity concerns and robustness to additional threats to validity later in the paper.

Our main finding is that brand-personality-related content drives social media engagement significantly, while directly informative content tends to drive engagement positively only when combined with such content. Additionally, directly informative content drives path-to-conversion (click-throughs). Combining both types of content thus enables the brand to obtain both the engagement and branding produced by brand personality-related content, as well as the immediate leads produced by directly informative content, along with any additional engagement they produce in combination. This finding is of substantial interest because most firms in our data post messages with one content type or other, rather than in combination. Our results suggest therefore that there may be substantial gains to content engineering by combining characteristics. Using only brand personality-related content does drive engagement, but using only this kind of content involves foregoing some of the benefits of obtaining leads and direct response. Similarly, using only direct informative content in posts is counterproductive because it reduces engagement, and thus reduces future reach due to the intermediating role of EdgeRank's filtering. This seems the main tradeoff between these two content types on the Facebook platform. Combining characteristics thus achieve a balanced tradeoff between reach and website visits. Our empirical results also unpack these effects into component attribute effects and also estimate the heterogeneity in these effects across firms and industries, enabling fine tuning these strategies across firms and industries.

Our paper adds to a growing literature on social media. Studies have examined the diffusion of user-generated content (Susarla et al., 2012) and their impact on firm performance (Rui et al., 2013; Dellarocas, 2006). A few recent papers have also examined the social media strategies of firms, focusing primarily on online blogs and forums. These include studies of the impacts of negative blog messages by employees on blog readership (Aggarwal et al., 2012), blog sentiment and quality on readership (Singh et al., 2014), social product features on consumer willingness to pay (Oestreicher-Singer and Zalmanson, 2013), and the role of active contributors on forum participation (Jabr et al., 2014). We add to this literature by examining the impact of firms' content strategies on user engagement.

An emerging theoretical literature in advertising has started to investigate the effects of content. This includes new models that allow ad content to matter in equilibrium by augmenting the canonical signaling model in a variety of ways (e.g. Anand and Shachar (2009)) by allowing ads to be noisy and targeted; Anderson and Renault (2006) by allowing ad content to resolve consumers' uncertainty about their match-value with a product; and Mayzlin and Shin (2011) and Gardete (2013) by allowing ad content to induce consumers to search for more information about a product. Our paper is most closely related to a small empirical literature that has investigated the effects of ad content in field settings. These include Bertrand et al. (2010) (effect of direct-mail ad content on loan demand); Anand and Shachar (2011); Liaukonyte et al. (2013) (effect of TV ad content on viewership and online sales); Tucker (2012a) (effect of ad persuasion on YouTube video sharing) and Tucker (2012b) (effect of "social" Facebook ads on philanthropic participation). Also related are recent studies exploring the effect of content more generally (and not specifically ad content) including Berger and Milkman (2012) (effect of emotional content in New York Times articles on article sharing) and Gentzkow and Shapiro (2010) (effect of newspaper's political content on readership). Relative to these literatures, our study makes two main contributions. First, from a managerial standpoint, we discuss the value of combining brand personality-related and directly informative content to balance reach/engagement with website clicks in social media, and demonstrate the differential effects of these types of content on consumer-oriented outcomes. This can help drive content engineering policies in firms. We also show how the effects differ by industry type. Second, none of the prior studies on ad content have been conducted at the scale of this study, which spans a large number of industries. We believe the content-tagging methodology we develop, which combines surveys implemented on AMT with NLP-based algorithms, provides a useful framework on which to build future studies that analyze the content of marketing communication.

We close this introduction with three caveats. First, we do not address the separate but important question of how engagement affects product demand and firm's sales so as to complete the link between ad-attributes and those outcome measures. The reader should note that the data required for the analysis of this question at a scale comparable to this study are still not widely available to researchers. Further, as mentioned, firms and advertisers care about engagement *per se* and are willing to invest in social media marketing for generating engagement, rather than caring only about sales. This is consistent with the view that advertising is a dynamic problem and a dominant role of advertising is to build long-term brand-capital for the firm. Even though the current period effects of advertising on demand may be small, the long-run effect of advertising may be large, generated by intermediary activities like increased consumer engagement, increased awareness and inclusion in the consumer consideration set. Thus, studying the formation and evolution of these intermediary activities − like engagement − is worthwhile in order to better understand the true mechanisms by which advertising affects outcomes in market settings. We further note that other papers such as Kumar et al. (2013); Goh et al. (2013); Rishika et al. (2013); Li and Wu (2014); Miller and Tucker (2013); Sunghun et al. (2014); Luo and Zhang (2013); Luo et al. (2013) as well as industry reports (comScore, 2013; Chadwick-Martin-Bailey, 2010; 90octane, 2012; HubSpot, 2013) have linked the social media engagement measures we consider to customer acquisition, sales, and profitability metrics.

Second, a caveat to our selection correction is that it is built on prior (but imperfect) knowledge of how EdgeRank is implemented. In the absence of additional experimental/exogenous variation, we are unable to address all possible issues with potential nonrandom assignment perfectly. We view our work as a large-scale exploratory study of content variables in social media that could be the basis of further rigorous testing and causal assessment, albeit at a more limited scale. A fully randomized large-scale experiment that provides a cross-firm and cross-industry assessment like provided here may be impossible or cost-prohibitive to implement, and hence, we think a large-scale cross-industry study based on field data of this sort is valuable.[2] Third, though we consider a larger range of content attributes than the existing literature, it is practically impossible to detail the full range of possible content profiles produced on a domain as large as Facebook (or in data as large as ours). We choose content profiles that reflect issues flagged in the existing academic literature and those that are widely used by companies on Facebook. We discuss this in more detail in Section 2.

2 Data

Our dataset is derived from the "pages" feature offered by Facebook. The feature was introduced on Facebook in November 2007. Facebook Pages enable companies to create profile pages and to post status updates, advertise new promotions, ask questions and push content directly to consumers. The left panel of Figure 1 shows an example of Walmart's Facebook Page, which is typical of the type of pages large companies host on the social network. In what follows, we use the terms pages, brands, and firms interchangeably. Our data comprises posts served from firms' pages onto the Facebook profiles of the users that are linked to the firm on the platform. To fix ideas, consider a typical message (see the right panel of Figure 1): "Pretty cool seeing Andy giving Monfils some love... Check out what the pros are wearing here: `http://bit.ly/nyiPeW`."[3] In this status update, a tennis equipment retailer starts with small talk, shares details about a celebrity (Andy Murray and Gael Monfils) and ends with link to a product page. Each such message is a unit of analysis in our data.

2.1 Data Description

2.1.1 Raw Data and Selection Criteria

To collect the data, we partnered with an anonymous firm, henceforth referred to as Company X that provides analytics services to Facebook Page owners by leveraging data from Facebook's *Insights*. *Insights* is

[2] A researcher that wishes to run an experiment to solve the problem would with luck, convince a subset of firms that host Facebook pages to run an experiment in which they randomly try different message types on the platform; however, conditional on this random allocation, whether or not the users actually obtains the posts in their newsfeed is always determined by Facebook's EdgeRank algorithm. This breaks randomization because the subset of treated users that EdgeRank chooses to show the post to is always a selected sample. The researcher or advertiser cannot solve this problem; only Facebook can. Thus, in our view, field experimentation at scale – though very desirable – is currently practically infeasible for the research community.

[3] Retailer picked randomly from an online search; not necessarily from our data.

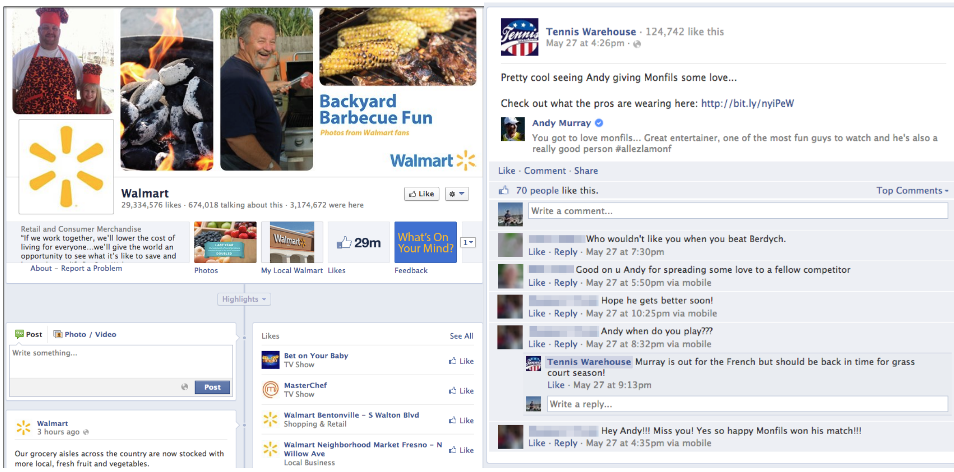

Figure 1: (Left) Example of a firm's Facebook Page (Walmart). (Right) Example of a firm's message and subsequent user engagement with that message (Tennis Warehouse). Example is not necessarily from our data.

a tool provided by Facebook that allows page owners to monitor the performance of their Facebook messages. Company X augments data from Facebook *Insights* across a large number of client firms with additional records of daily message characteristics, to produce a raw dataset comprising a message-day-level panel of messages posted by companies via their Facebook pages. The data also includes two consumer engagement metrics: the number of *Likes* and comments for each message each day. These metrics are commonly used in industry as measures of engagement. They are also more granular than other metrics used in extant research such as the number of fans who have *Liked* the page. Also available in the data are the number of impressions of each message per day (i.e., the total number of users the message is exposed to). In addition, page-day level information such as the aggregate demographics of users (fans) who *Liked* the page on Facebook or have ever seen messages by the page are collected by Company X on a daily level. This comprises the population of users a message from a firm can potentially be served to. We leverage this information in the methodology we develop later for accounting for non-random assignment of messages to users by Facebook. Once a firm serves a message, the message's impressions, *Likes,* and comments are recorded daily for an average of about 30 days (maximum: 126 days).[4] The raw data contains about a million unique messages by about 2,600 unique companies.

The reader should note that to the best of our knowledge, as of this writing our data is the most complete observational data available outside of Facebook – the data includes details such as demographics of page fans and engaged fans, which cannot be scraped by outsiders (but are essential for correcting for *EdgeRank*) but are available only to the page owners via Facebook's Application Programming Interface. Our data also includes daily snapshots of message-level engagement that Facebook provides to page owners. These daily snapshots generate the within-message variation that enables the panel analysis in our paper. Finally, page-owners do not have access to data on performance of any messages by other pages, unlike our dataset

[4]The majority of messages do not get any impressions or engagement after 7 days. After 15 days, virtually all engagement and impressions (more than 99.9%) are accounted for.

which spans a large number of companies across sectors.

We clean the data to reflect the following criteria: (i) only pages located in the US, (ii) only messages written in English, and (iii) only messages with complete demographic data. After cleaning, the data span 106,316 unique messages posted by 782 companies (including many large brands) between September 2011 and July 2012. This results in about 1.3 million rows of message-level daily snapshots recording about 450 million page fans' responses. Removing periods after which no significant activity is observed for a message reduces this to 665,916 rows of message-level snapshots (where activity is defined as either impressions, *Likes*, or comments). The companies in our dataset are categorized into 6 broader industry categories following Facebook's page classification criteria: Celebrities & Public Figure (e.g., Roger Federer), Entertainment (e.g., Star Trek), Consumer Products & Brands (e.g., Tesla Motors), Organizations & Company (e.g., WHO), Websites (e.g., TED), Local Places & Businesses (e.g., MoMA).

2.1.2 Content-coded Data

We use a two-step method to label content. First, we contract with workers through AMT and tag 5,000 messages for a variety of content profiles. Subsequently, we build an NLP algorithm by combining several statistical classifiers and rule-based algorithms to extend the content-coding to the full set of 100,000 messages. This algorithm uses the 5,000 AMT-tagged messages as the training data-set. We describe these methods in more detail later in the paper.

Table 2 outlines the finer classification of the attributes we code up, including precise definitions, summary statistics, and the source for coding the attribute. We broadly group the messages as directly informative, brand personality-related, or both. Some messages inform consumers about deals and discounts about products, while other messages seek to connect with consumers on a personal level to promote brand personality, form relationships and are social in nature. We call the first type directly informative content, and the second brand personality-related content. Some messages do both at the same time by including casual banter and product information simultaneously (e.g., "Are you a tea person or a coffee person? Get your favorite beverage from our website: http://www.specific-link-here.com").

Directly informative content variables are identified using the work of Resnik and Stern (1977), which provides fourteen evaluative content criteria to identify directly informative content that includes content such as product price, deals and availability. In Table 2, the 8 variables: BRANDMENTION, DEAL, PRICECOMPARE, PRICE, TARGET, PRODAVAIL, PRODLOCATION, and PRODMENTION are directly informative. These variables enable us to assess the effect of search attributes, brand, price, and product availability information on engagement. Our brand personality-related content are chosen by picking attributes that occur commonly in Facebook's posts and also reflect those pointed out as important in driving consumer response in existing consumer behavior research. For example, emotional and humorous content have been identified as drivers of virality (Porter and Golan, 2006; Berger, 2012, 2011; Berger and Milkman, 2012). Philanthropic content has been studied in the context of advertising effectiveness (Tucker,

Sample Messages	**Content Tags**
Maria's mission is helping veterans and their families find employment. Like this and watch Maria's story. http://walmarturl.com/VzWFlh	PHILANTHROPIC, SMALLTALK, ASKLIKE, HTTP
Cheers! Let Welch's help ring in the New Year.	BRANDMENTION, SMALLTALK, HOLIDAYMENTION, EMOTION
On a scale from 1-10 how great was your Christmas?	SMALLTALK, QUESTION, HOLIDAYMENTION
Score an iPad 3 for an iPad2 price! Now at your local store, $50 off the iPad 3. Plus, get a $30 iTunes Gift Card. Offer good through 12/31 or while supplies last.	PRODMENTION, DEAL, PRODLOCATION, PRODAVAIL, PRICE

Table 1: **Examples of Messages and Their Content Tags:** The messages are taken from 2012 December messages on Walmart's Facebook page.

2012b). Similarly, Berger and Schwartz (2011) documented that the interestingness of content such as mentions of remarkable facts is effective in generating word-of-mouth. The 8 variables: REMFACT, EMOTION, EMOTICON, HOLIDAYMENTION, HUMOR, PHILANTHROPIC, FRIENDLIKELY, and SMALLTALK are classified as brand personality-related. These definitions include emotional content, humor, banter, and philanthropic content. While not fully exhaustive, we have attempted to cover most variables that are 1) highlighted by prior academic research to be relevant, 2) commonly discussed and used in the industry.

Besides these main variables of interest, controls and content-related patterns noted as important in industry reports are profiled. We include these content categories to investigate more formally considerations laid out in industry white papers, trade-press articles, and blogs about the efficacy of message attributes in social media engagement. It includes content that explicitly solicits readers to comment or includes blanks for users to fill out (thus providing an explicit option to facilitate engagement). Additionally, characteristics like whether the message contained photos, website links, and the types of the page-owner (e.g., business organization versus celebrity) are also coded. Other message-specific characteristics and controls include metrics such as message length in characters and SMOG ("Simple Measure of Gobbledygook"), an automatically computed reading complexity index that is used widely. Higher values of SMOG implies a message is harder to read. Table 1 shows sample messages taken from Walmart's page in December 2012 and shows how we would have tagged them. The reader should note that some elements of content tagging and classification are necessarily subjective and based on human judgement. We discuss our methods (which involve obtaining agreement across 9 tagging individuals) in section 2.2. All things considered, we believe this is one of the most comprehensive attempts at tagging marketing communication related content in the empirical literature.

2.1.3 Data Descriptive Graphics

This section presents descriptive statistics of the main stylized patterns in the data. While there is active interest in social media, very little is known (even at a descriptive level) about what kinds of content are commonly used by firms. Hence, we first report on what kinds of content are used by firms. Table 2

Variable	Description	Source	Mean	SD	Min	Max
TAU (τ)	Time since the post release (Day)	Facebook	6.253	3.657	1	16
LIKES	Number of "*Likes*" post has obtained	Facebook	48.373	1017	0	324543
COMMENTS	Number of "Comments" post has obtained	Facebook	4.465	78.19	0	22522
IMPRESSIONS	Number of times message was shown to users (unique)	Facebook	9969.2	129874	1	4.5×10^7
SMOG	SMOG readability index (higher means harder to read)	Computed	7.362	2.991	3	25.5
MSGLEN	Message length in characters	Computed	157.41	134.54	1	6510
HTTP	Message contains a link	Computed	0.353	0.478	0	1
QUESTION	Message contains questions	Computed	0.358	0.479	0	1
BLANK	Message contains blanks (e.g. "My favorite artist is _ _")	Computed	0.010	0.099	0	1
ASKLIKE	Explicit solicitation for "*Likes*" (e.g. "Like if ...")	Computed	0.006	0.080	0	1
ASKCOMMENT	Explicit solicitation for "Comments"	Computed	0.001	0.029	0	1
MSGTYPE	Categorical message type assigned by the Facebook	Facebook				
- App	application related messages	Facebook	0.099	0.299	0	1
- Link	link	Facebook	0.389	0.487	0	1
- Photo	photo	Facebook	0.366	0.481	0	1
- Status Update	regular status update	Facebook	0.140	0.347	0	1
- Video	video	Facebook	0.005	0.070	0	1
PAGECATEGORY	Page category closely following Facebook's categorization	Facebook				
- Celebrity	Singers, Actors, Athletes etc	Facebook	0.056	0.230	0	1
- ConsumerProduct	consumer electronics, packaged goods etc	Facebook	0.296	0.456	0	1
- Entertainment	Tv shows, movies etc	Facebook	0.278	0.447	0	1
- Organization	non-profit organization, government, school organization	Facebook	0.211	0.407	0	1
- PlaceBusiness	local places and businesses	Facebook	0.071	0.257	0	1
- Website	page about a website	Facebook	0.088	0.283	0	1
	Brand personality-related					
REMFACT	Remarkable fact mentioned	AMT	0.527	0.499	0	1
EMOTION	Any type of emotion present	AMT	0.524	0.499	0	1
EMOTICON	Contains emoticon or net slang (approximately 1000 scraped from web emoticon dictionary e.g. :D, LOL)	Computed	0.012	0.108	0	1
HOLIDAYMENTION	Mentions US Holidays	Computed	0.006	0.076	0	1
HUMOR	Humor used	AMT	0.375	0.484	0	1
PHILANTHROPIC	Philanthropic or activist message	AMT	0.498	0.500	0	1
FRIENDLIKELY	Answer to question: "Are your friends on social media likely to post message such as the shown"?	AMT	0.533	0.499	0	1
SMALLTALK	Contains small talk or banter (defined to be content other than about a product or company business)	AMT	0.852	0.355	0	1
	Directly informative					
BRANDMENTION	Mentions a specific brand or organization name	AMT+Comp	0.264	0.441	0	1
DEAL	Contains deals: any type of discounts and freebies	AMT	0.620	0.485	0	1
PRICECOMPARE	Compares price or makes price match guarantee	AMT	0.442	0.497	0	1
PRICE	Contains product price	AMT+Comp	0.051	0.220	0	1
TARGET	Message is targeted towards an audience segment (e.g. demographics, certain qualifications such as "Moms")	AMT	0.530	0.499	0	1
PRODAVAIL	Contains information on product availability (e.g. stock and release dates)	AMT	0.557	0.497	0	1
PRODLOCATION	Contains information on where to obtain product (e.g. link or physical location)	AMT	0.690	0.463	0	1
PRODMENTION	Specific product has been mentioned	AMT+Comp	0.146	0.353	0	1

Table 2: **Variable Descriptions and Summary for Content-coded Data:** To interpret the "Source" column, note that "Facebook" means the values are obtained from Facebook, "AMT" means the values are obtained from Amazon Mechanical Turk and "Computed" means it has been either calculated or identified using online databases and rule-based methods in which specific phrases or content (e.g. brands) are matched. Finally, "AMT+Computed" means primary data has been obtained from Amazon Mechanical Turk and it has been further augmented with online resources and rule-based methods.

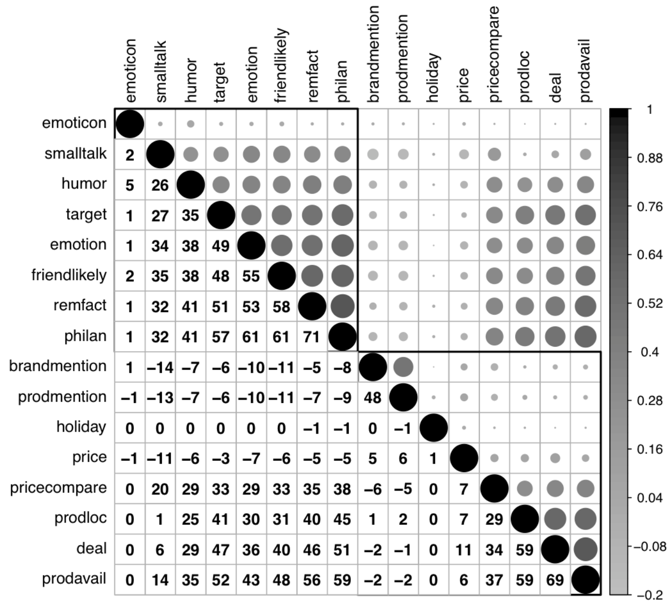

Figure 2: **Co-occurrence of Attribute Characteristics Across messages**. Shades in upper triangle represent correlations. Numbers in lower triangle represent the same correlations in numerical form in 100-s of units (range -100,+100). For e.g., the correlation in occurrence of humor and smalltalk across messages is 0.26 (cell [3,2]). The dark line shows the separation into 2 clusters. Brand personality-related content and directly informative content attributes tend to form two separate clusters.

reports the mean proportion of messages that have each content characteristic. One can see that messages with videos, price information, holiday mentions or emoticons are relatively uncommon, while those with smalltalk and with information about where to obtain the product (PRODAVAIL/PRODLOCATION) are very common. Figure 2 reports on the co-occurrence of the various attributes across messages. The patterns are intuitive. For instance, emotional and philanthropic content co-occur often, so does emotional and friend-like content, as well as content that describes product deals and availability. To better describe the correlation matrix graphically and to cluster highly correlated variables together, we ran cluster analysis (hierarchical clustering with the number of clusters determined using the average silhouette width (Rousseeuw, 1987)), which suggested that there are two clusters in the data. Figure 2 shows via a solid line how content types are clustered across messages. We see that brand personality-related content types and directly informative content types are roughly split into two separate clusters, suggesting that firms typically tend to use one or the other in their messages. Later in the paper, we show evidence suggesting that this strategy may not be optimal.

Figure 3 shows the percentage of messages featuring a content attribute split by industry category. We represent the relative percentages in each cell by the size of the bubbles in the chart. The largest bubble is SMALLTALK for the celebrities category (60.4%) while the smallest is PRICECOMPARE for the celebrities category (0%). This means that 6 in 10 messages by celebrity pages in the data have some sort of small talk (banter) and/or content that does not relate to products or brands; and that there are no messages by celebrity owned pages that feature price comparisons. "Remarkable facts" (our definition) are posted more by firms in the entertainment category and less by local places and businesses. Consistent with intuition, consumer product pages and local places/businesses post the most about products (PRODMENTION), product availability (PRODAVAIL), product location (PRODLOC), and deals (DEAL). Emotional (EMOTION) and philanthropic (PHILAN) content have high representation in pages classified as celebrity, organization, and websites.

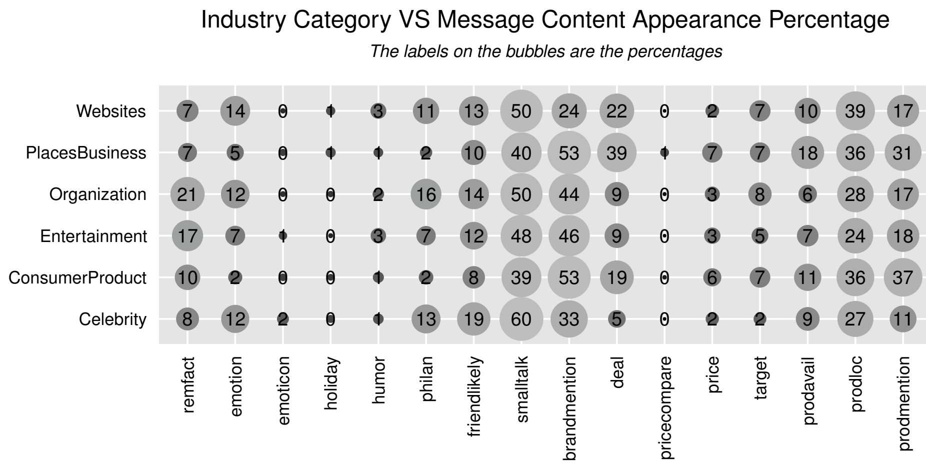

Figure 3: **Bubble Chart of Broader Industry Category vs Message Content:** Each bubble represents the percentage of messages within a row-industry that has the column-attribute. Computed for the 5000 tagged messages. Larger and lighter bubbles imply higher percentage of messages in that cell. Percentages do not add up to 100 along rows or columns as any given message can have multiple attributes included in it. The largest bubble (60.4%) corresponds to SMALLTALK for the celebrity page category and the smallest bubble (0%) corresponds to PRICECOMPARE for the celebrity category.

We now discuss the engagement data. Figure 4 shows box plots of the log of impressions, *Likes*, and comments versus the time (in days) since a message is released (τ). Both comments and *Likes* taper off to zero after two and six days respectively. The rate of decay of impressions is slower. Virtually all engagements and impressions (more than 99.9%) are accounted for within 15 days of release of a message.

Figure 5 shows the average number of *Likes* and comments by message type (photo, link, etc.) over the lifetime of a message. Messages with photos have the highest average *Likes* (94.7) and comments (7.0) over their lifetime. Status updates obtain more comments (5.5) on average than videos (4.6) but obtain less *Likes* than videos. Links obtain the lowest *Likes* on average (19.8) as well as the lowest comments (2.2). Figure 6 shows the same bar plots split across 6 industry categories. A consistent pattern is that messages with photos always obtain highest *Likes* across industries. The figure also documents interesting heterogeneity in engagement response across industries. The patterns in these plots echo those described in reports by many market research companies such as Wildfire and comScore.

Figure 7 presents the average number of *Likes* and comments by content attribute. Emotional messages obtain the most number of *Likes* followed by messages identified as "likely to be posted by friends" (variable: FRIENDLIKELY). Emotional content also obtain the highest number of comments on average followed by

SMALLTALK and FRIENDLIKELY. The reader should note these graphs do not account for the market-size (i.e., the number of impressions a message reached). Later, we present an econometric model that incorporates market-size as well as selection by Facebook's filtering algorithm to assess user engagement.

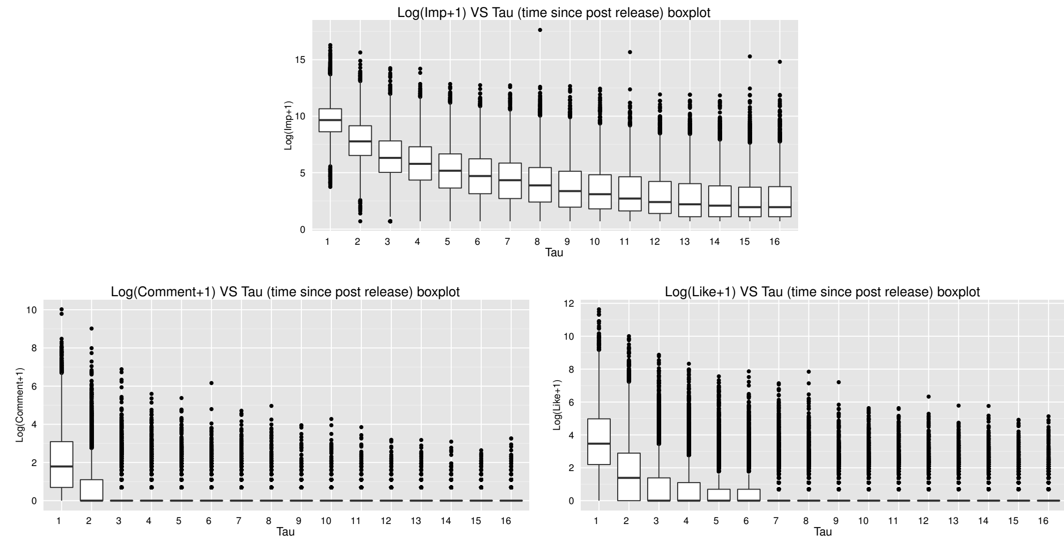

Figure 4: **: Box Plots of Log(engagement+1) vs Time since message Release:** Three graphs show the box plots of (log) impressions, comments and *Like* vs. τ respectively. Both comments and *Likes* taper to zero after two and six days respectively. Impressions take longer. After 15 days, virtually all engagement and impressions (more than 99.9%) are accounted for. There are many outliers.

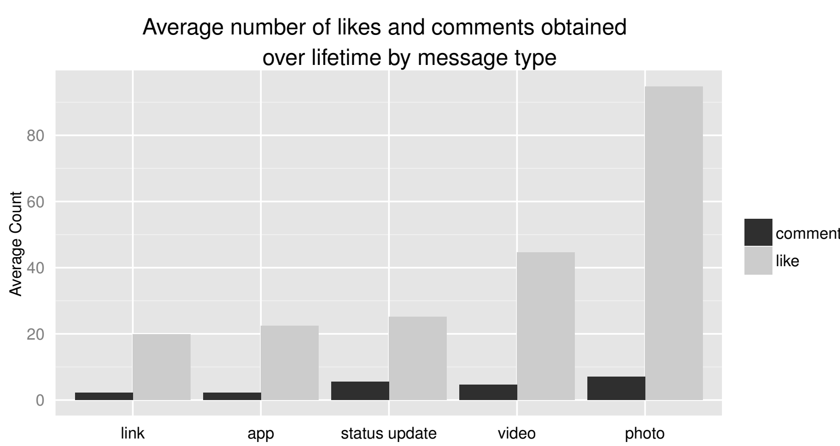

Figure 5: **Average *Likes* and Comments by Message Type**: This figure shows the average number of *Likes* and comments obtained by messages over their lifetime on Facebook, split by message type.

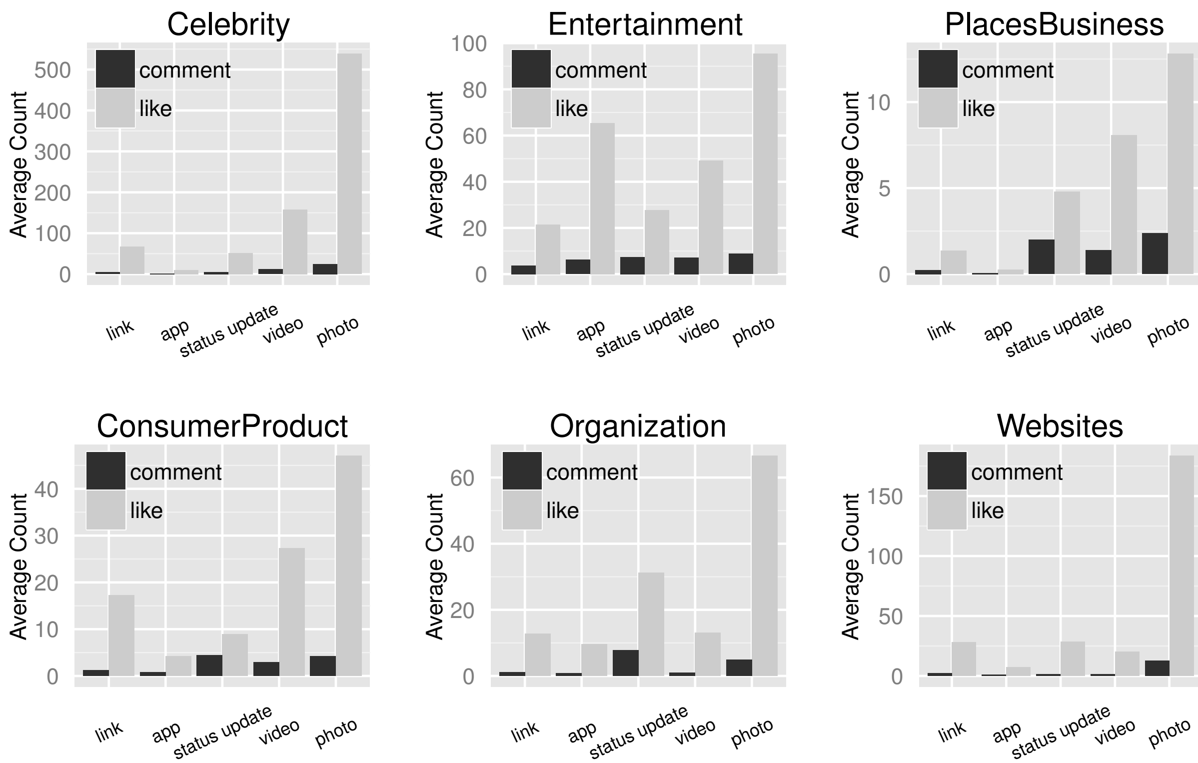

Figure 6: **Average *Likes* and Comments by Message Type by Industry**: This figure shows the average number of *Likes* and comments obtained by messages over their lifetime split by message type for each industry.

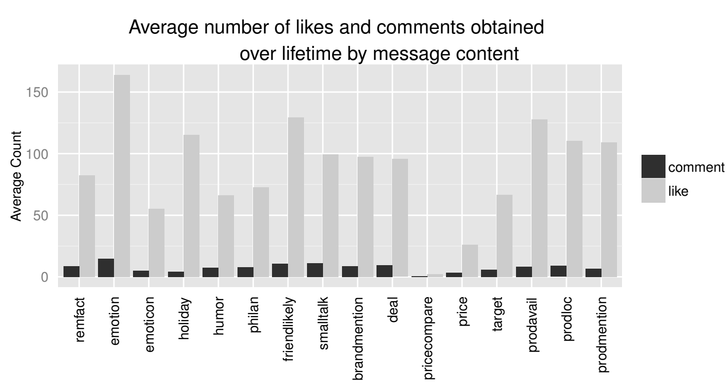

Figure 7: **Average *Likes* and Comments by Message Content**:This figure shows the average number of *Likes* and comments obtained by messages over their lifetime split by message content.

2.2 Amazon Mechanical Turk (AMT)

We now describe our methodology for content-coding messages using AMT. AMT is a crowd sourcing marketplace for simple tasks such as data collection, surveys, and text analysis. It has now been successfully leveraged in several academic papers for online data collection and classification. To content-code our mes-

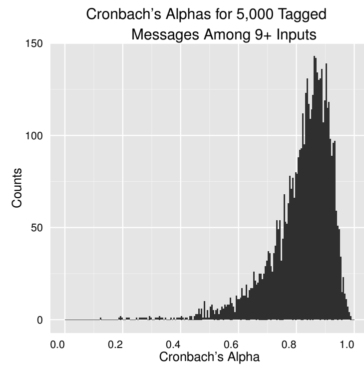

Figure 8: **Cronbach's Alphas for 5,000 Messages**: This bar graph shows the inter-rater reliability measure of Cronbach's Alpha among at least 9 distinct Turkers' inputs for each 5,000 messages. The mean is 0.82 and the median is 0.84. We replicated the study with only those above 0.7 and found the result to be robust.

sages, we create a survey instrument comprising of a set of binary yes/no questions we pose to workers (or "Turkers") on AMT. To ensure high quality responses from the Turkers, we follow several best practices identified in literature (e.g., we obtain tags from at least 9 different Turkers choosing only those who are from the U.S., have more than 100 completed tasks, and an approval rate more than 97%. We also include an attention-verification question.) Please see the appendix for the final survey instrument and the complete list of strategies implemented to ensure output quality.

Figure 8 presents the histogram of Cronbach's Alphas, a commonly used inter-rater reliability measure, obtained for the $5,000$ messages.[5] The average Cronbach's Alpha for our $5,000$ tagged messages is 0.82 (median 0.84), well above typically acceptable thresholds of 0.7. About 87.5% of the messages obtained an alpha higher than 0.7, and 95.4% higher than 0.6. For robustness, we replicated the study with only those messages with alphas above 0.7 (4,378 messages) and found that our results are qualitatively similar.

At the end of the AMT step, approximately $2,500$ distinct Turkers contributed to content-coding $5,000$ messages. This constitutes the training dataset for the NLP algorithm used in the next step.

2.3 Natural Language Processing (NLP) for Attribute Tagging

We use NLP techniques to label message content from Facebook messages using the AMT-labeled messages as the training data. Typical steps for such labeling tasks include: 1) breaking the sentence into understandable building blocks (e.g., words or lemmas) and identifying sentence-attributes similar to what humans do when reading; 2) obtaining a set of training sentences with labels tagged from a trusted source identifying whether the sentences do or do not have a given content profile (in our case, this source comprise the 5000 AMT-tagged messages); 3) using statistical tools to infer which sentence-attributes are correlated with content outcomes, thereby learning to identify content in sentences. When presented with a new set of sentences,

[5]Recall, there are at least 9 Turkers per message. We calculate a Cronbach's Alpha for each message by computing the reliability across the 9 Turkers, across all the content classification tasks associated with the message. Figure 8 then plots a histogram of the Cronbach's Alphas so computed across the 5,000 messages.

the algorithm breaks the sentence down to building blocks, identifies sentence-level attributes, and assigns labels using the statistical models that were fine-tuned in the training process. We summarize our method here briefly. A detailed description of the algorithms employed is presented in the Appendix.

The use of NLP methods has gained traction recently in business research due to the readily available text data online (e.g., Netzer et al. (2012); Ghose et al. (2012); Geva and Zahavi (2013)). Our NLP methods closely mirror large-scale, multi-step methods used in the financial services industry to automatically extract financial information from textual sources (e.g., Hassan et al. (2011)) and are similar in flavor to winning algorithms from the recent Netflix Prize competition.[6] Our method combines five statistical classifiers with rule-based methods via heterogeneous "ensemble learning" to build-up a final, master classifier. The *statistical classifiers* we use are essentially binary classification machine learning models that take attributes as input and output predicted classification probabilities. We fit a variety of classifiers to our training dataset. These include logistic regression with L1 regularization (which penalizes the number of attributes and is commonly used for attribute selection for problems with many attributes; see (Hastie et al., 2009)), Naive Bayes (a probabilistic classifier that applies Bayes theorem based on presence or absence of features), and support vector machines (a gold-standard algorithm in machine learning that works well for high dimensional problems) with L1 and L2 regularization and various kernels including linear, radial basis function, and polynomial kernels. We also utilize class-weighted classifiers and resampling methods to account for imbalance in positive and negative labels. The *rule-based methods* we use are essentially algorithms that use large data sources (a.k.a dictionaries) or use specific i*f-then* rules inputted by human experts, to scan through particular words or occurrences of linguistic entities in the messages to generate a classification. We use a variety of rule-based methods. For example, in identifying brand and product mentions, we augment our AMT-tagged answers with several large lists of brands and products from online sources and a company list database from Thomson Reuters. Further, to increase the range of our brand name and product database, we also ran a separate AMT study with 20,000 messages in which we asked AMT Turkers to identify any brand or product name included in the message. We added all the brand and product names we harvested this way to our look-up database. We then utilize a set of rules to identify brand and product mentions by looking up these lists. Similarly, in identifying emoticons in the messages, we use large dictionaries of text-based emoticons freely available on the internet.

Finally, we combine the classifications from the many classifiers and rule-based algorithms we use together via ensemble learning methods. Combining classifiers this way has several advantages since a single statistical classifier cannot successfully overcome the classical precision-recall tradeoff inherent in the classification problem.[7] The final combined classifier has higher precision and recall than any of the constituent classifiers. It is important to note that our algorithm is optimized to identify specific content in 106,000 messages in

[6]See `http://www.netflixprize.com`.

[7]The performance of NLP algorithms are typically assessed on the basis of accuracy (the total % correctly classified), precision (out of predicted positives, how many are actually positive), and recall (out of actual positives, how many are predicted as positives). An important tradeoff in such algorithms is that an increase in precision often causes decrease in recall or vice versa. This tradeoff is similar to the standard bias-variance tradeoff in estimation.

	With Ensemble Learning (The Best Performing Algorithm)			Without Ensemble Learning (Support Vector Machine version 1 + Rule-based)		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
REMFACT	0.94	0.99	0.68	0.88	0.99	0.33
EMOTION	0.97	0.99	0.87	0.94	0.98	0.65
HUMOR	0.98	1	0.90	0.97	1	0.14
PHILANTHROPIC	0.97	0.99	0.85	0.93	0.99	0.62
FRIENDLIKELY	0.94	0.99	0.68	0.90	0.99	0.41
SMALLTALK	0.85	0.88	0.80	0.78	0.34	0.28
DEAL	0.94	0.99	0.65	0.90	1	0.43
PRICECOMPARE	0.99	0.99	1	0.99	1	0.85
TARGETING	0.98	0.99	0.89	0.95	0.99	0.71
PRODAVAILABILITY	0.96	0.99	0.76	0.91	1	0.10
PRODLOCATION	0.97	0.99	0.90	0.87	1	0.11

Table 3: **Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation**: This table presents metrics for performance of the classification algorithms used. The left 3 columns show the metrics for the final algorithm which combines classifiers via ensemble learning methods while the right 3 columns shows the metrics for a support vector machine algorithm. Notice that the support vector machine classifier tends to have low recall and high precision. Naive Bayes tends to have high recall but low precision. Classifiers on their own cannot successfully overcome the standard precision-recall tradeoff (if one is higher, the other is lower). But combining many different classifiers with ensemble learning can increase both precision and recall. We obtain similar results for negative class labels.

our dataset and not general content in any text.

Assessment We assess the performance of the overall NLP algorithm on three measures, viz., accuracy, precision, and recall (as defined in Footnote 7) using 10-fold cross-validation. 10-fold cross-validation is computationally intensive and makes it harder to achieve higher accuracy, precision and recall in general. However, we find using the 10-fold criterion critical to obtaining the external validity required for large scale classification. Table 3 shows these metrics for different content profiles. The performance is extremely good and comparable to performance achieved by the leading financial information text mining systems (Hassan et al., 2011). We also report the improvement of the final ensemble learning method relative to using only a support vector machine classifier. As shown, the gains from combining classifiers are very substantial. We obtain similar results for negative class labels.

As a final point of assessment, note that several papers in the management sciences using NLP methods implement *unsupervised* learning which does not require human-tagged data. These techniques use existing databases such as WordNet (lexical database for English) or tagged text corpus (e.g, tagged Brown Corpus) to learn content by patterns and correlations. *Supervised* NLP instead utilizes human-taggers to obtain a robust set of data that can be used to train the algorithm by examples. While unsupervised NLP is inexpensive, its performance is relatively poor compared to that of supervised NLP algorithms like the ones implemented here. Finally, to the best of our knowledge, the NLP method used in this paper that uses ensemble learning to combine several statistical classifiers and rule-based methods, has not been used in business research journals. Further, several current implementations of NLP do not utilize the strict bar of utilizing the 10-fold cross-validation criterion. We believe one of the contributions of this paper is to demonstrate how to utilize AMT in combination with ensemble learning techniques, to implement supervised

NLP in business research to produce robust and cost-efficient NLP algorithms that perform well at the scale required for empirical work. We believe the method will be useful in future studies on unstructured natural language data such as advertising content or product reviews. For interested readers, a detailed step-by-step description of our NLP algorithm's training and classification procedures is presented in the Appendix.

3 Empirical Strategy

Our empirical goal is to investigate the effect of message ad content on subsequent customer engagement. Engagement − the y-variable − is observed in the data; and content − the x-variables − has been tagged as above and is also observed. If messages are randomly allocated to users, the issue of assessing the effect of message-content on engagement is straightforward; one simply projects y on x. Unfortunately, a complication arises because Facebook's policy of delivery of messages to users is non-random: users more likely to find a message appealing are more likely to see the message in their newsfeed, a filtering implemented via Facebook's "EdgeRank" algorithm. The filtering implies a selection problem in estimation of the effect of message-characteristics on engagement − if we see that messages with photos are more likely to be commented on by users, we do not know if this is due to consumer propensity to comment on messages with photos, or whether Facebook is very effective in showing messages with photos to users who are more likely to comment on them. To our knowledge, the issue has been ignored in the literature on social media analysis so far.[8] We address the selection issue via a two-step procedure, first by building a semi-parametric model of EdgeRank that delivers an estimate of the expected number of impressions a message is likely to receive, and then, by incorporating this model to run a selectivity-corrected projection of *Likes* and comments on message characteristics in the second-stage. For the first-stage, we exploit the fact that we observe the aggregated decisions of Facebook to serve impressions to users, and that EdgeRank is based on three variables as revealed by Facebook: Type, Tie, and Time.[9]

- *Type* (z) refers to the type of message. Facebook categorizes message-type into 5 classes: status update, photo, video, app, or link.
- *Tie* (h_{ijt}) refers to the affinity score between page j (company) and the Facebook user i (viewer of the message) at time t which is based on the strength and frequency of the interaction history between the user and the page.
- *Time* (τ) refers to the time since the message.

Our dataset contains direct observations on the variables Type and Time. We do not have individual-level data on a user's history with pages to model tie strengths. However, we exploit the fact that we have access

[8] We discuss later in the paper why other sources of confounds (like direct targeting by firms) are second-order in this setting, compared to the selection induced by EdgeRank-based filtering. Section (4.4) presents some sensitivity analysis and robustness to assess these other sources of potential endogeneity bias.

[9] As disclosed first at the 2010 "f8" conference. See `http://whatisEdgeRank.com` for a brief description of EdgeRank. For the duration of our data collection, this EdgeRank specification holds true.

to demographics data on the set of users who *could potentially have been shown* a message, versus *who were actually shown* the message. The difference reflects the selection by EdgeRank, which we utilize as a proxy measure of Tie-strength based targeting. Since we do not know the exact functional form of EdgeRank's targeting rule, we work with a semi-parametric specification, utilizing flexible splines to capture the effect of EdgeRank. At the end of this step, we thus develop a flexible approximation to EdgeRank's targeting. In the second step, we can then measure the effect of ad content on *Likes* and comments, by controlling for the non-random targeting using our first-stage model. Figure 9 shows the empirical strategy visually. The advantages of directly modeling EdgeRank this way, are that 1) by separating Facebook's impression mechanism from the effect of content on consumer engagement, our results speak to consumer behavior and not Facebook's filtering algorithm, thereby increasing external validity of our results to realms outside of Facebook Pages and 2) we are also able to predict which message would eventually reach users in addition to handing selection, which has auxiliary managerial value for advertisers seeking higher reach.

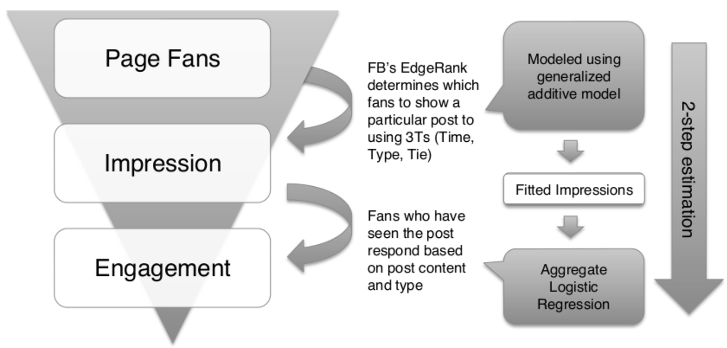

Figure 9: **Impression-Engagement Funnel:** Facebook's EdgeRank chooses subset of Page fans to show messages released by the page and fans who've seen the message engage with the message based on content and type. EdgeRank is modeled with a semi-parametric model (generalized additive model) and the final engagement is estimated through aggregate logistic regression. Details of estimation are in Sections 3.1 and 3.2.

3.1 First-stage: Approximating EdgeRank's Assignment

We represent message k's type in a vector z_k, the time since message k was released in τ_k, and the history of user i's past engagement with company j on Facebook in a vector h_{ijt}. Table 4 summarizes the notation.

To understand our procedure, let $n_{kjt}^{(d)}$ denote the number of users of demographic type $d = 1, .., D$ who were shown message k by firm j at time t. We refer to $n_{kjt}^{(d)}$ as impressions. We observe n_{kjt} directly, and $n_{kjt}^{(d)}$ is indirectly reported in the data and can be reverse-engineered from Company X's reports. A description of this procedure is provided in appendix. Let $\mathbb{N}_{jt}^{(d)}$ denote the total number of users of demographic type d for firm j on day t to whom the message can potentially be delivered. $\mathbb{N}_{jt}^{(d)}$ is directly observed in the data, and comprises all users of demographics d who have *Liked* the firm on Facebook. To be clear, note that *Liking* a message is different from *Liking* a page $-$ *Liking* a page provides the firm that maintains that page an opportunity to serve its messages to that user via Facebook's Newsfeed. $\mathbb{N}_{jt}^{(d)}$ is a count of all such users.

If posts were randomly served to users by Facebook, then the demographic distribution of $n_{kjt}^{(d)}$ would be identical to the distribution of $\mathbb{N}_{jt}^{(d)}$. The difference is due to EdgeRank. Now, note that by EdgeRank's assignment rule, the aggregated impressions for demographic type d, $n_{kjt}^{(d)}$, is an (unknown) function of liked-fans $\mathbb{N}_{jt}^{(d)}$, the tie strength between users within demographic bucket d and the posting firm, $h_{ijt}^{(d)}$, the type of message z_k, and time since message release τ_k,

$$\mathbb{E}(n_{kjt}^{(d)}) = g(\mathbb{N}_{jt}^{(d)}, h_{ijt}^{(d)}, z_k, \tau_k) \tag{1}$$

We do not observe individual-level data on each users i's interaction with every message which could be the basis of estimating Equation (1). Instead, we can construct the aggregated number of impressions and liked-fans within a set of demographic buckets in the data. To use this variation as a source of approximating EdgeRank, we approximate the RHS of Equation (1) as,

$$\mathbb{E}(n_{kjt}^{(d)}) \approx g_d(\mathbb{N}_{jt}^{(d)}, \theta_{1j}^{(d)}, z_k, \tau_k) \tag{2}$$

where, we use a firm-demographic bin specific fixed effect, $\theta_{1j}^{(d)}$, to capture the effect of user history. This approximation would literally be true if all individuals within demographic bucket d had the same history with firm j. In practice, this is not the case, and this may induce approximation errors into the procedure, because additional history-heterogeneity within demographic buckets is not modeled (or is assumed into the error term). This is a caveat to our analysis. Access to individual-level data could be the basis of improving this procedure and relaxing this assumption. We view Equation (2) as a flexible approximation that allows us to leverage the observed variation in firm-level impressions across demographics, while enabling us to include firm and demographic-level fixed effects into a procedure that best approximates EdgeRank based on what we as researchers (and firms) know about Facebook's filtering algorithm. We will also estimate the right-hand function $g_d(.)$ separately for each demographic bucket, in effect allowing for slope heterogeneity in demographics in addition to intercept heterogeneity across demographics.

The next step relates to approximating the function $g_d(.)$. Since we do not know the exact functional form

Notation	***Description***
i	User
j	Firm
k	message
t	Time (day)
z_k	message k's media type (5 options: photo, video, status update, app, link)
τ_k	Time since message k was released
h_{ijt}	History of user i's past engagement with firm j
$g(.)$	EdgeRank score approximating function
$n_{kjt}^{(d)}$	Impressions of message k by page j at time t by users in demographics bin d
$\mathbb{N}_{jt}^{(d)}$	Number of users of demographics bin d who *Liked* page j as of time t
$\theta_0^{(d)}$	Intercept term for each demographics d
$\theta_{\cdot}^{(d)}$	Parameters in EdgeRank approximation for demographics bin d

Table 4: **User-level Setup Notation**

of the above selection equation, we approximate the function semi-parametrically via a Generalized Additive Model (GAM) (c.f., Hastie and Tibshirani (1990)). The GAM is a generalized linear model with additive predictors consisting of smoothed (e.g. interpolation and curve fitting) covariates. It provides increased flexibility in approximating the unknown function, $g_d(.)$. The GAM fits the following flexible relationship between a set of covariates X and dependent variable Y,

$$\mu(\mathbb{E}(Y|X_1, X_2, ..., X_p)) = \alpha + s_1(X_1) + s_2(X_2) + ... + s_p(X_p)$$

where μ is a link function (e.g. gaussian, poisson, gamma), and $s_1, s_2, ...s_p$ are nonparametric smoothing functions such as cubic splines or kernel smoothers. We model the EdgeRank selection equation for each demographic d as the following,

$$\begin{aligned} h_d\left[\log(n_{kjt}^{(d)}+1)\right] &= \theta_0^{(d)} + \theta_{1j}^{(d)} + \theta_2^{(d)}\mathbb{N}_{jt}^{(d)} + s_1(\mathbb{N}_{jt}^{(d)};\theta_3^{(d)}) + \sum_{r=2}^{5}\theta_{4r}^{(d)}\mathbb{I}\left(z_k = r\right) \\ &+ \sum_{r=2}^{16}\theta_{5r}^{(d)} I\left(\tau_k = r\right) + \epsilon_{kjt}^{(d)} \end{aligned} \tag{3}$$

where, $h_d \equiv g_d^{-1}(.)$ is the identity (Gaussian) link function, $\theta_0^{(d)}$ is an intercept term unique to each demographic, d, and $\theta_{1j}^{(d)}$ is a firm-demographic fixed effect that captures the tie strength between the firm j and demographics d.[10] $\mathbb{N}_{jt}^{(d)}$ is the number of fans of demographic d for firm j at time t and denotes the potential audience for a message. s_1 is a cubic spline smoothing function, essentially a piecewise-defined function consisting of many cubic polynomials joined together at regular intervals of the domain such that the fitted curve, the first and second derivatives are continuous. We represent the interpolating function $s_1(.)$ as a linear combination of a set of basis functions $b(.)$ and write: $s_1(\mathbb{N}_{jt}^{(d)};\theta_3^{(d)}) = \sum_{r=3}^{q} b_r\left(\mathbb{N}_{jt}^{(d)}\right)\theta_{3r}^{(d)}$, where the $b_r(.)$ are a set of basis functions of dimension q to be chosen and $\theta_{3.}^{(d)}$ are a set of parameters to be estimated. We follow a standard method of generating basis functions, $b_r(.)$, for the cubic spline interpolation as defined in Wood (2006). Fitting the spline also requires choosing a smoothing parameter, which we tune via generalized cross-validation. We fit all models via the R package `mgcv` described in Wood (2006).

Finally, we include dummy variables for message-type (z_k) and for each day since release of the message (τ_k; up to 16 days), to capture the effect of message-type and time-since-release semi-parametrically. These are allowed to be $d-$specific. We collect the set of parameters to be estimated for each demographic bucket in a vector, $\theta_{.}^{(d)}$, which we estimate by GAM estimation. The estimated parameter vector, denoted $\hat{\theta}_{.}^{(d)}, d = 1, .., D$, serves as an input to the second stage of the estimation procedure.

[10] We also tried Poisson and Negative Binomial link functions (since $n_{kjt}^{(d)}$ is a count variable), as well as the identity link function without logging the y-variable. Across these specifications, we found the identity link function with $\log(y)$ resulted in the best fit, possibly due to many outliers. We also considered specifications with numerous interaction of the covariates included, but found they were either not significant or provided trivial gains in the R^2. Lastly, removing the extreme outliers did not change the results qualitatively.

3.2 Second-stage: Modeling Engagement Given Message Assignment

We operationalize engagement via two actions, *Likes* and comments on the message. The selection problem is that users can choose to *Like* or comment on a message only if they are served impressions, which generates non-random censoring because impression assignment is endogenous to the action. We address the censoring by including a correction for the fact that a user is shown a message non-randomly, estimated semi-parametrically as above. Suppose $\hat{\Psi}_{kjt}^{(d)}$ denotes the fitted estimate from the first-stage of the expected number of impressions of message k for firm j amongst users of type d at time t,

$$\hat{\Psi}_{kjt}^{(d)} = g_d\left(N_{jt}^{(d)}, z_k, \tau_k; \hat{\theta}^{(d)}\right) \tag{4}$$

We model the probability that users of type-d will *Like* a message given the full set of message characteristics, M_{kt}, as logistic with parameters $\Omega = (\delta_d, \psi)_{d=1..D}$,

$$\pi_d(M_{kt}; \Omega) = \frac{1}{1 + e^{-(\delta_d + M_{kt}\psi)}} \tag{5}$$

The parameter vector, Ω, is the object of inference in the second stage.[11]

We will estimate Ω by fitting the model to explain Q_{kjt}, the observed number of *Likes* of the message in each period in the data. To see the intuition for how our correction works in the estimation, note that we can aggregate Equation (5) across users, so that the expected number of *Likes* is,

$$\mathbb{E}(Q_{kjt}; \Omega) \approx \sum_{d=1}^{D} \hat{\Psi}_{kjt}^{(d)} \times \left[\frac{1}{1 + e^{-(\delta_d + M_{kt}\psi)}}\right] \tag{6}$$

with $\hat{\Psi}_{kjt}^{(d)}$ are treated as known from the first-stage (Equation 4). The right-hand side is a weighted sum of logit probabilities of *Liking* a message. Intuitively, the decision to *Like* a message is observed by the researcher only for a subset of users who were endogenously assigned an impression by FB. The selection functions $\hat{\Psi}_{kjt}^{(d)}$ serve as weights that reweigh the probability of *Liking* to account for the fact that those users were endogenously sampled, thereby correcting for the non-random nature of message assignment when estimating the outcome equation.

We could use the expectation in Equation (6) as the basis of an estimation equation. Instead, for efficiency, we estimate the parameter vector Ω by maximum likelihood. To set up the likelihood, note the expected number of impressions of message k for firm j at time t across *all* demographic buckets is simply the sum,

$$\hat{\Psi}_{kjt} = \sum_{d=1}^{D} g_d\left(N_{jt}^{(d)}, z_k, \tau_k; \hat{\theta^{(d)}}\right) \tag{7}$$

We can obtain an estimate of the implied probability that an impression picked at random from the pool is

[11] Allowing ψ to be d-specific as well in Equation (5) is conceptually straightforward. Unfortunately, this results in parameter proliferation and trouble with convergence; hence we settled for a more limited specification with d-specific intercepts.

of type-d,

$$\hat{\varrho}_{dkt} = \frac{\hat{\Psi}_{kjt}^{(d)}}{\hat{\Psi}_{kjt}} \tag{8}$$

Thus, the probability $\pi(M_{kt};\Omega)$ that an impression picked at random from the pool will *Like* the message given a guess of Ω, is,

$$\pi(M_{kt};\Omega) = \sum_{d=1}^{D} \mathbb{P}_{kt}(d) \times \mathbb{P}_{kt}(Like|d) = \sum_{d=1}^{D} \hat{\varrho}_{dkt} \times \pi_d(M_{kt};\Omega) \tag{9}$$

Intuitively, with probability $\mathbb{P}_{kt}(d) = \hat{\varrho}_{dkt}$ an impression is of type-d, and with probability $\mathbb{P}(Like|d) = \pi_d(M_{kt};\Omega)$, an impression will *Like* the message conditional on being type-d; hence the unconditional probability a random impression will *Like* the message is the sum-product of these marginals and conditionals across all D types.

The number of *Likes* is a count variable for which we specify a Binomial likelihood. Accordingly, the probability that Q_{kjt} out of the $\hat{\Psi}_{kjt}$ assigned impressions are observed to *Like* the message, and that $\hat{\Psi}_{kjt} - Q_{kjt}$ of the remaining impressions are observed not to, is binomial with probability, $\pi(M_{kt};\Omega)$,

$$Q_{kjt} \sim \text{Binomial}(\hat{\Psi}_{kjt}, \pi(M_{kt};\Omega)) \tag{10}$$

Maximizing the implied binomial likelihood across all the data, treating $\hat{\Psi}_{kjt}$ as given, then delivers estimates of Ω. The intuition for the selection correction here is the same as that encapsulated in Equation (6). We can repeat the same procedure using the number of comments on the message as the dependent variable so as the recover the effect of message-characteristics on commenting as well. This two-step procedure thus delivers estimates of the effects of message-characteristics on the two outcomes of interest. Standard errors are obtained by bootstrapping both steps 1 and 2 over the entire dataset.

Discussion The approach we have outlined above essentially uses the EdgeRank approximation as a control function (Heckman and Robb (1986)) that corrects for the selectivity in the second stage, where we measure the effect of message characteristics on outcomes. Intuitively, we exploit the observed discrepancy in demographic distributions between the set of individuals to whom a message could have been served, $N_{jt}^{(d)}$, versus those who were actually served, $n_{kjt}^{(d)}$. The discrepancy reflects the filtering by EdgeRank. Our first stage essentially projects this discrepancy onto message-type, time-since-release, page and demographic characteristics in a flexible way. Our control function relies on partial knowledge of the assignment/selection rule, and develops a model of how the non-random assignment of posts to users is implemented on the platform. If we knew EdgeRank perfectly, this would be the efficient (and preferred) solution to the selection problem. Since we know EdgeRank only partially, we are worried about misspecification of the control function. The flexible semi-parametric first-stage along with the inclusion of page-demographic specific fixed effects mollifies concerns about bias from this misspecification. We also tried several different alternative specifications with different link functions (including linear, poisson and negative binomial) obtaining qualitatively similar results, but inferior fit corresponding to our preferred specifications. Below we also show that the results we obtain from the EdgeRank approximation pass several sanity checks such as time-since-post having a negative effect on the probability of a post being served by EdgeRank, and the fact that time-coefficients monotonically decline as time-since-release of the post increases as reported in several industry studies from comScore/Wildfire. The demographic-page fixed effects, which correspond to the demographic-page "affinity" level, also coincide with expected patterns (e.g., the demographic bin that has the highest affinity with the newborn clothing page is the one corresponding to male and females 25-34, no significant effects for older demographics – discussed below). Notwithstanding these aspects, to the best of our knowledge, the full details of EdgeRank are not known to any firm or researcher. In our view, a "perfect" solution to the selection problem is unlikely to be achieved without full knowledge of Facebook's targeting rule. For approaches that have a similar flavor, please see Manchanda et al. (2004); Nair et al. (2013) in the context of targeted marketing; Ellickson and Misra (2010) in the context of selectivity correction in an entry game; and in particular, Ahn and Powell (1993) for semi/nonparametric control function approaches.

4 Results

4.1 First-Stage

The first-stage model, as specified in Equation 3, approximates EdgeRank's message assignment algorithm. We run the model separately for each of the 14 age-gender bins used by Facebook. These correspond to two gender and seven age bins. For a given bin, the model relates the number of users of demographic type d who were shown message k by firm j at time t to the message type (z_k), days since message (τ), and tie between the firm and the user. Table 5 presents the results. The intercepts ($\theta_0^{(d)}$) indicate that messages by companies in our dataset are shown most often to Females ages 35-44, Females 45-54, and Males 25-34. The lowest number of impressions are for the 65+ age group. In our model, tie between a user and a firm is proxied by a fixed-effect for each firm-demographic pair. This implies 800 $\times$ 14 fixed effects corresponding to 800 firms and 14 demographic bins. Due to space constraints, we do not present all the estimated coefficients. Table 5 presents the coefficients for two randomly chosen firms. The first is a new-born clothing brand and the second is a protein bar brand. For ease of visualization, these fixed effects are shown graphically in Figure 10 (only the statistically significant coefficients are plotted). For messages by the the new-born clothing brand, the most impressions are among females in the age-groups of 25-34, 18-24, and 35-44. Among males, ages 25-34 receive the most number of impressions. For messages by the protein bar brand, impressions are more evenly distributed across the different demographic bins, with the Male 18-24 group receiving the most impressions. These estimated coefficients are consistent with our expectations for the two brands.

The estimates for message type are roughly the same in all demographic bins. For all demographics, the photo type has the highest coefficient (around 0.25) suggesting that photos are preferred to all other media types by EdgeRank. This is likely because users have historically engaged better with photos causing Facebook to show photos more often. The next most preferred message type is the status update with coefficients averaging around 0.12 followed by videos and links. The baseline message type, apps, is the message type that is least preferred by EdgeRank. The rank ordering of coefficients for message type do not strictly follow the rank ordering of number of messages released by firms, which is shown in Table 2. Whereas links are posted more often, photos get more impressions relative to messages of other types, clearly highlighting the role of EdgeRank. Days since message (τ) are not presented in Table 5 due to space constraints. However, Figure 11 presents a box plot of the coefficients for τ across all 14 demographic bins. All coefficients are negative and significant and also more negative for higher values of τ, implying that EdgeRank prefers to show more recent messages. Finally, the coefficients for number of fans, $\mathbb{N}_{jt}^{(d)}$, are positive and significant but they have relatively low magnitude. This is because our model includes a smoothed term of the number of fans, $s(\mathbb{N}_{jt}^{(d)})$, which soaks up both the magnitude and nonlinearity. The smoothed fan-numbers are all significant.

	Female						
	F 13-17	**F 18-24**	**F 25-34**	**F 35-44**	**F 45-54**	**F 55-64**	**F 65+**
Intercept	5.528***	6.071***	6.446***	7.165***	7.209***	6.133***	4.887***
Page 1 fixed effect - new born clothing brand	-0.210	2.458***	2.685***	1.544**	0.888	0.813	0.489
Page 2 fixed effect - protein bar brand	-0.573***	1.285***	1.466***	0.928***	0.016	1.671***	1.518***
Message Type - App is the base							
Link	0.010	0.045***	0.063***	0.042***	0.051***	0.051***	0.048***
Photo	0.253***	0.318***	0.340***	0.309***	0.297***	0.267***	0.249***
Status Update	0.100***	0.161***	0.175***	0.152***	0.152***	0.129***	0.114***
Video	0.033	0.041	0.061**	0.041	0.021	0.024	0.030
$\mathbb{N}_{jt}^{(d)}$ (Fan Number)	2.0×10^{-6}***	1.8×10^{-6}***	7.2×10^{-6}***	1.9×10^{-5}***	1.9×10^{-5}***	3.8×10^{-5}***	8.5×10^{-5}***
$s(\mathbb{N}_{jt}^{(d)})$ significance	***	***	***	***	***	***	***
R-Squared	0.78	0.78	0.77	0.78	0.78	0.78	0.77
	Male						
	M 13-17	**M 18-24**	**M 25-34**	**M 35-44**	**M 45-54**	**M 55-64**	**M 65+**
Intercept	5.486***	6.118***	7.075***	6.635***	6.125***	5.151***	4.011***
Page 1 fixed effect - new born clothing brand	0.156	0.932	1.673**	1.082	0.722	0.209	0.111
Page 2 fixed effect - protein bar brand	1.867***	2.423***	0.907***	0.670***	1.158***	1.575***	1.502***
Message Type - App is the base							
Link	-0.005	0.025***	0.033***	0.034***	0038***	0.049***	0.030***
Photo	0.226***	0.284***	0.295***	0.277***	0.254***	0.230***	0.212***
Status Update	0.077***	0.124***	0.126***	0.120***	0.106***	0.103***	0.084***
Video	0.014	0.039	0.044*	0.031	0.016	0.007	0.023
$\mathbb{N}_{jt}^{(d)}$ (Fan Number)	3.6×10^{-6}***	1.0×10^{-6}***	6.7×10^{-6}***	2.5×10^{-5}***	3.8×10^{-5}***	5.2×10^{-5}***	2.3×10^{-4}***
$s(\mathbb{N}_{jt}^{(d)})$ significance	***	***	***	***	***	***	***
R-Squared	0.79	0.80	0.79	0.78	0.78	0.77	0.76
*App is the base for message type. Significance Level: '***' <0.001 '**' < 0.01 '*' 0.05							

Table 5: **EdgeRank Model Estimates**: This table presents the coefficients obtained from 14 generalized additive models for EdgeRank, calculated for each demographic bin. There are 14 demographic (gender-age) bins provided by Facebook. F13-17 means all females in the age between 13 and 17. Time since message (τ), and page-level fixed effects are not included in the table and presented graphically separately.

The generalized additive model of EdgeRank recovers coefficients that make intuitive sense and are consistent with claims made in several industry reports (e.g. that photos have the highest EdgeRank weight). Further, the model fit appears to be good especially given that we have used generalized cross-validation to guard against overfitting.

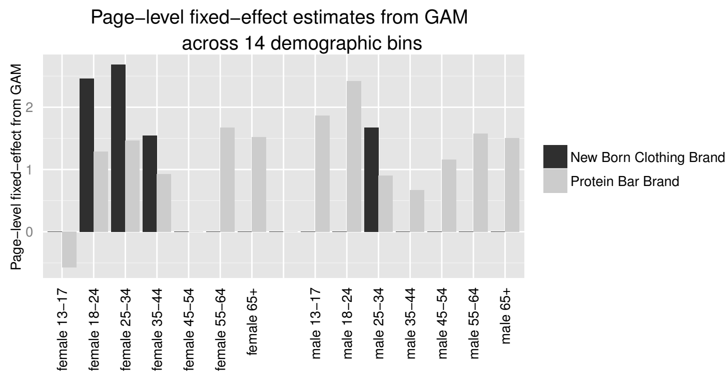

Figure 10: **Page-level Fixed effect Estimates from Generalized Additive Model Across 14 Demographic Bins:** This bar graph shows two randomly chosen page-level fixed effect estimates from the EdgeRank models. Only the statistically significant estimates are shown. New born clothing brands are positively significant for 18-24 female, 25-34 female, 35-44 female, and 25-34 male. Protein bar brands have the highest fixed effect among 18-24 male demographics.

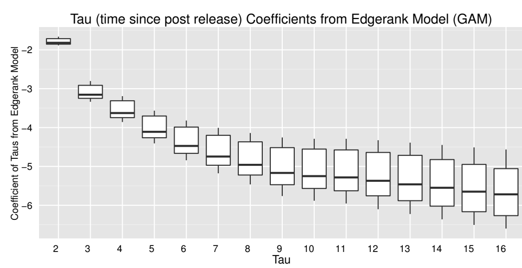

Figure 11: **Time Since message Release (τ) Coefficients Box plot Across Demographics:** This box plot shows the coefficients on τ across all the demographics bin. $\tau = 1$ is the base case and every coefficients are significant at the highest level of $p < 0.001$.

4.2 Second-Stage

In the second-stage, we measure the effect of content characteristics on engagement using our selectivity-corrected model from the first-stage. All results in this section are based on an analysis of the entire set of over 100,000 messages (i.e. the 5,000 AMT-tagged messages as well as the messages tagged using NLP). The results for only the 5,000 AMT-tagged messages are qualitatively similar and are presented in the appendix. To present the results in a simple way, we first create the two composite summary variables corresponding to brand personality-related content and directly informative content. The brand personality-related variable is obtained by adding values of REMFACT, EMOTION, EMOTICON, HOLIDAYMENTION, HUMOR, PHILANTHROPIC, FRIENDLIKELY, and SMALLTALK resulting in a composite variable ranging from

0 to 8. The directly informative composite variable is obtained by adding values of BRANDMENTION, DEAL, PRICECOMPARE, PRICE, TARGET, PRODAVAIL, PRODLOCATION, and PRODMENTION resulting in a composite variable ranging from 0 to 8. Table 6 shows the result of logistic regression on engagement with these composite variables and interaction of those two variables as the x-s.

We find that inclusion of more brand personality-related content has a positive and statistically significant effect on both types of engagement; further, inclusion of more directly informative content reduces engagement. Interestingly, the interaction between brand personality-related and directly informative content is positive, implying that directly informative content increases engagement in the presence of brand personality-related content in the message. What could explain this pattern of results? One possible reason why engagement decreases as directly informative content increases is that too much of informative content all at once may be interpreted as direct selling, which spoils the experience of consuming the platform, analogous to the way advertisement generally spoilt the TV viewing experience. Another reason may be the disconnect that such content produces with the user's frame of mind when logging into Facebook. For instance, it may be jarring to be served ads about discounts, low prices and sales when a user logs into the platform for social interaction and for checking updates from friends and family. Another possible reason is that that users don't engage publicly – through visible Likes, Comments and Shares–with directly informative content but respond directly by clicking on relevant informative posts by clicking on links and visiting the website. This may again be due to the disconnect between directly informative content and the rest of the social content on Facebook. These are merely conjectures. Investigating the underling mechanism is beyond the scope of the current data; nevertheless, the sign of the effect is robust across specifications. Finally, the improved engagement associated with brand personality-related content may be driven by the congruence of the brand personality with consumers' own personalities and because the benefits of a brand become more persuasive when expressed by the brand's personality, as suggested in the introduction. Such content seems to help firms in relationship-building and to persuade consumers to engage with their brand via such relationships.

Variable	Comment	Like
Constant	-6.913***(0.002)	-4.671***(0.001)
Brand personality-related	0.053***(0.001)	0.061***(0.000)
Directly informative	-0.143***(0.001)	-0.068***(0.000)
Brand personality-related × Directly informative	0.012***(0.000)	0.003***(0.000)
McFadden R-sq.	0.015	0.009
Nagelkerke R-sq.	0.015	0.009
Log-likelihood	-4208220.431	-33678695.014
Deviance	8012471.987	66409947.187
AIC	8416448.861	67357398.028
N	665916	665916
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1		

Table 6: **Brand personality-related vs Directly informative:** Logistic regression for {Comment, Like} with composite summary variables for brand personality-related and directly informative content.

Table 7 presents the results of aggregate logistic regression with the full list of content variables. We present results for both engagement metrics (*Likes*/comments) as well as for models with and without the EdgeRank correction. We exclude the 16 estimated τ coefficients from the table since they are all negative and statistically significant just as in the EdgeRank model in Figure 11. We also exclude demographic fixed effects for space.

	NO ER COMMENT	OR	ER COMMENT	OR	NO ER LIKE	OR	ER LIKE	OR
Constant	12.309*** (0.197)		14.083*** (0.142)		-7.383*** (0.089)		13.504*** (0.065)	
SMOG	-0.045*** (0.000)	0.956	-0.066*** (0.000)	0.936	-0.029*** (0.000)	0.971	-0.057*** (0.000)	0.945
MSGLEN	0.000 (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000
HTTP	-0.484*** (0.002)	0.616	-0.324*** (0.002)	0.723	-0.353*** (0.000)	0.703	-0.180*** (0.000)	0.835
QUESTION	0.449*** (0.001)	1.567	0.527*** (0.001)	1.694	-0.292*** (0.000)	0.747	-0.185*** (0.000)	0.831
BLANK	0.942*** (0.003)	2.565	1.099*** (0.003)	3.001	-0.716*** (0.002)	0.489	-0.625*** (0.002)	0.535
ASKLIKE	0.002 (0.010)	1.002	0.178*** (0.010)	1.195	0.456*** (0.003)	1.578	0.501*** (0.003)	1.650
ASKCOMMENT	0.779*** (0.021)	2.179	0.710*** (0.021)	2.034	-0.090*** (0.011)	0.914	-0.282*** (0.011)	0.754
Brand personality-related								
REMFACT	-0.019*** (0.002)	0.981	0.010*** (0.002)	1.010	-0.060*** (0.001)	0.942	-0.035*** (0.001)	0.966
EMOTION	0.203*** (0.002)	1.225	0.257*** (0.002)	1.293	0.201*** (0.001)	1.223	0.257*** (0.001)	1.293
EMOTICON	0.118*** (0.004)	1.125	-0.053*** (0.004)	0.948	-0.132*** (0.001)	0.876	-0.214*** (0.001)	0.807
HOLIDAYMENTION	-0.493*** (0.014)	0.611	-0.352*** (0.014)	0.703	-0.323*** (0.004)	0.724	-0.136*** (0.004)	0.873
HUMOR	0.023*** (0.002)	1.023	0.082*** (0.002)	1.085	-0.044*** (0.000)	0.957	0.012*** (0.000)	1.012
PHILANTHROPIC	0.147*** (0.002)	1.158	0.140*** (0.002)	1.150	0.008*** (0.001)	1.008	0.001 (0.001)	1.001
FRIENDLIKELY	0.002 (0.002)	1.002	-0.022*** (0.002)	0.978	0.073*** (0.001)	1.076	0.058*** (0.001)	1.060
SMALLTALK	0.045*** (0.002)	1.046	-0.074*** (0.002)	0.929	-0.052*** (0.001)	0.949	-0.121*** (0.001)	0.886
Directly informative								
BRANDMENTION	0.000 (0.002)	1.000	0.077*** (0.002)	1.080	-0.031*** (0.000)	0.969	0.004*** (0.000)	1.004
DEAL	-0.163*** (0.002)	0.850	-0.168*** (0.002)	0.845	-0.198*** (0.001)	0.820	-0.200*** (0.001)	0.819
PRICECOMPARE	-0.031*** (0.001)	0.969	0.005*** (0.001)	1.005	-0.037*** (0.000)	0.964	-0.040*** (0.000)	0.961
PRICE	-0.104*** (0.005)	0.901	-0.319*** (0.005)	0.727	-0.187*** (0.001)	0.829	-0.400*** (0.001)	0.670
TARGET	-0.016*** (0.002)	0.984	-0.073*** (0.002)	0.930	0.030*** (0.001)	1.030	-0.041*** (0.001)	0.960
PRODAVAIL	-0.067*** (0.002)	0.935	-0.060*** (0.002)	0.942	-0.109*** (0.001)	0.897	-0.065*** (0.001)	0.937
PRODLOCATION	-0.054*** (0.002)	0.947	0.009*** (0.002)	1.009	0.063*** (0.001)	1.065	0.107*** (0.001)	1.113
PRODMENTION	-0.050*** (0.002)	0.951	-0.148*** (0.002)	0.862	0.077*** (0.001)	1.080	-0.007*** (0.001)	0.993
Message Type - App is the base								
-Link	0.177*** (0.003)	1.194	-0.238*** (0.003)	0.788	0.126*** (0.001)	1.134	-0.374*** (0.001)	0.688
-Photo	0.867*** (0.003)	2.380	0.519*** (0.003)	1.680	1.011*** (0.001)	2.748	0.651*** (0.001)	1.917
-Status Update	1.146*** (0.003)	3.146	0.818*** (0.003)	2.266	0.478*** (0.001)	1.613	0.060*** (0.001)	1.062
-Video	-0.106*** (0.009)	0.899	0.466*** (0.009)	1.594	-0.200*** (0.003)	0.819	0.341*** (0.003)	1.406
Industry Category - Celebrity is the base								
-ConsumerProduct	0.171*** (0.002)	1.186	-0.319*** (0.002)	0.727	-0.369*** (0.001)	0.691	-0.813*** (0.001)	0.444
-Entertainment	0.362*** (0.002)	1.436	0.418*** (0.002)	1.519	-0.291*** (0.001)	0.748	-0.256*** (0.001)	0.774
-Organization	0.485*** (0.002)	1.624	0.281*** (0.002)	1.324	-0.004*** (0.001)	0.996	-0.212*** (0.001)	0.809
-PlaceBusiness	0.429*** (0.005)	1.536	0.021*** (0.005)	1.021	-0.639*** (0.002)	0.528	-1.109*** (0.002)	0.330
-Websites	0.012*** (0.003)	1.012	0.074*** (0.003)	1.077	0.088*** (0.001)	1.092	0.118*** (0.001)	1.125
McFadden R-sq.	0.271		0.207		0.32		0.239	
Nagelkerke R-sq.	0.271		0.207		0.321		0.241	
Log-likelihood	-2,446,467.133		-3,423,466.377		-14,108,100.91		-25,950,910.53	
Deviance	4,488,295.547		6,443,162.73		27,268,539.27		50,955,992.81	
AIC	4,893,058.266		6,847,056.753		28,216,325.82		51,901,945.06	
N	665,916		665,916		665,916		665,916	
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1								

Table 7: **Aggregate Logistic Regression Results For Comments and *Likes***: This table presents the aggregate logistic regression on comments and *Likes* for both EdgeRank-corrected (ER) and uncorrected (NO ER) for all data. OR means Odds ratio and shows the odds ratio for the estimates left of the column.

Scanning through the results, we observe that the estimates are directionally similar, in most cases, with and without EdgeRank correction. However, the magnitudes often change. For example, consider the coefficients for message type Photo. In the model without EdgeRank correction, Photos are very likely to get comments (coefficient = 0.867) and *Likes* (coefficient = 1.011). After EdgeRank correction, the results are similar but the magnitude of the effect drops. This makes sense because we know that EdgeRank prefers Photos. In some instances, there are directional changes for some coefficients. For example, the result that links are more likely to get *Likes*/comments relative to apps changes sign after EdgeRank correction. This highlights the importance of EdgeRank correction, an issue that most industry reports (e.g., Wildfire 2012) often overlook.

We find that high reading complexity (SMOG) decreases both *Likes* and comments whereas shorter messages (MSGLEN) are *Liked* and commented on more, albeit with a small effect size. Having links (HTTP) is associated with lower engagement whereas asking questions (QUESTION) significantly increase comments but at the cost of *Likes*. Using blanks in the message to encourage comments has a similar effect of increasing comments but hurting *Likes*. Interestingly, while the odds ratio of comments increases by 69% if a message asks a question, it increases by 200% if blanks are included suggesting that blanks are more effective than questions if the goal is to increase comments. Asking for *Likes* increase both *Likes* and comments, whereas asking for comments increase comments but at the cost of *Likes*. It is clear that even these simple content variables impact user engagement.

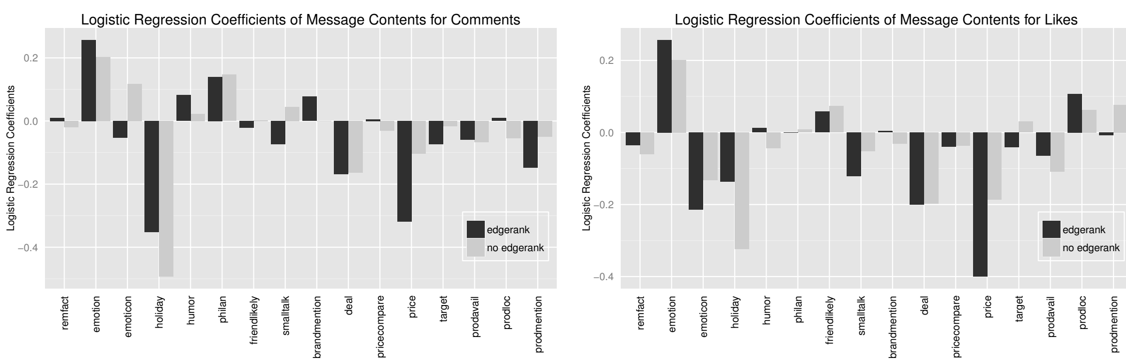

Figure 12: **Message Characteristic Coefficients for Comments and *Likes***: These bar graphs show the coefficients of logistic regression for both EdgeRank corrected and uncorrected models. Only the significant coefficients are plotted.

The next 16 variables in the table are the brand personality-related and directly informative content variables. Figure 12 charts the coefficients for these variables in a bar graph and demonstrates the sharp difference between brand personality-related and directly informative content types. Looking at comments, a striking pattern is that most directly informative contents have a negative impact whereas brand personality-related contents have a positive impact. The directly informative content variables with the most negative impact are PRICE, DEAL, and PRODMENTION. The brand personality-related content variables with the most positive impact are EMOTION and PHILANTHROPIC. Interestingly, HOLIDAYMENTION discourages comments. One possible explanation is that near holidays, all Facebook pages indiscriminately mention holidays, leading to dulled responses. For example, during Easter, the occurrence of holiday mention jumped to nearly 40% across all messages released that day compared to the average occurrence of about 1%. Looking at *Likes*, fewer brand personality-related content variables have positive impact but the results are qualitatively similar to that for comments. Among brand personality-related contents, EMOTION has the most positive impact on *Likes*. Most directly informative content variables continue to have a negative impact (i.e., reduce engagement), with PRICE and DEAL having the most negative impact. The results also highlight that there exist differences between the impact of content on *Likes* versus Comments.

We also investigate how industry moderate the impact of content on engagement, we repeated the main analysis on subsetted data on each industry type as described and categorized in Table 2. Some coefficients are different across industries both in magnitude and, for some variables, in direction. However, there are also many coefficients that are persistent across the industry. Additionally, comparing Figures 3 and 17 provides interesting comparisons of what each industry is currently posting and what users engage with. Due to the length of the paper, we provide more discussion in Appendix.

4.3 Shares, Click-throughs, and Additional Exploration of Results

The above results suggest that brand personality-related content increases engagement while directly informative content in general reduces it (unless combined with brand personality-related content). Does this imply that firms do not benefit from using directly informative content? This conclusion would be incorrect as directly informative content could drive other profitable behaviors other than engagement. For instance, consumers who appreciate the economic/utilitarian aspect of informative, direct-response messages may directly proceed to purchase the product or take other conversion actions, which are beneficial to the firm. To the extent that we observe only engagement and not conversion/sales data, these outcomes are missed by the current analysis. While it would be ideal to augment our analysis with sales data, our data partner does not have data on purchase data and most firms do not track such information at the post level. Nevertheless, to assess these, we augment our data with two additional pieces of information. These additional data include:

1. Total cumulative *clicks* from the date of release till the last date of the data on each of the 106,000 messages in our database.
2. Total cumulative *shares* from the date of release till the last date of the data on each of the 106,000 messages in our database.

These clicks would likely be the first step in a series of actions leading to eventual purchase. Hence, they are an important piece of the puzzle to explore the direct response effects of informative messages. The shares data are another important engagement variable on Facebook and enable us to check the robustness of our findings for Likes/Comments.[12]

[12] We thank the journal's review team for encouraging us to explore these ideas further. Unfortunately, we were unable to obtain ultimate sales/conversion information. Our data partner does not have data on purchase activity occurring on the third-party websites of its clients. This data reside only with the individual firms that run the websites.

We then run the full series of models in the paper again using clicks and shares as dependent variables. Figure 13 shows the results. The figure depicts the coefficients on the content attributes from running the model with shares and clicks respectively as the dependent variables. Looking at the shares model, we see the basic findings from the other engagement models are replicated – attributes with brand personality-related content like emotion and humor have positive effects on sharing the post, while direct information like mentions of deals and prices and price comparisons in particular do not obtain many shares from users.

The results for clicks are also presented below. In contrast to the previous results, we find the effect of deal information is now highly positive. While presenting information about deals (i.e. discounts) does not seem to elicit Likes, comments, and shares, we find evidence they increase click-through rates. The results for other content attributes are qualitatively similar to those for Likes and comments. The other attribute that has a positive effect on click-through is the mention of a holiday in the post (likely reflecting the presence of discounts and deals at the firm's website during holidays). Both suggest that such information, while reducing engagement, may set the consumer on a path to conversion. These results suggest a more nuanced interpretation of the effects of brand personality-related versus directly informative content. Brand personality-related content primarily drives engagement and seems key for long term brand building, while directly informative content drives direct-response and seems key to performance marketing.

Content Design These results then imply guidelines for better content design. Our results suggest content design should be driven by the tradeoffs between the goals of the campaign for the firm. We see the trade-off between directly informative and brand personality-related content as one between immediate leads (click-throughs) versus future visibility in the social media site and branding (from engagement). Since brand personality-related content drives engagement and certain directly informative content drive path-to-conversion and do not reduce engagement when combined with brand personality-related content, it seems combining both together, when feasible, would form the basis of improved content engineering. Using only brand personality-related content does drive engagement, but using only this kind of content involves foregoing some of the benefits of obtaining website traffic and direct response. Similarly, using direct informative content in messages helps the firm gain direct leads, but to repeat only this kind of content in post after post may be counterproductive. Such content will not only have poor social reach but, since the EdgeRank algorithm uses a firm's current engagement to determine the future reach of the firm's posts, repeatedly posting information-exclusive content will eventually diminish the firm's future reach. Thus, while direct informative messages help facilitate clicks and potential conversion, the difficulty is that these occur at the cost of reducing the size of future reached-fan-base. This seems the main tradeoff between these two content types on the Facebook platform. Combining characteristics thus achieves a balanced tradeoff between reach and potential conversion. We documented earlier in section (2.1.3) that firms typically tend to use one content-type or the other in their messages. Our current results suggest this strategy may not be optimal that the gains to improved content management along the lines suggested above may be substantial.

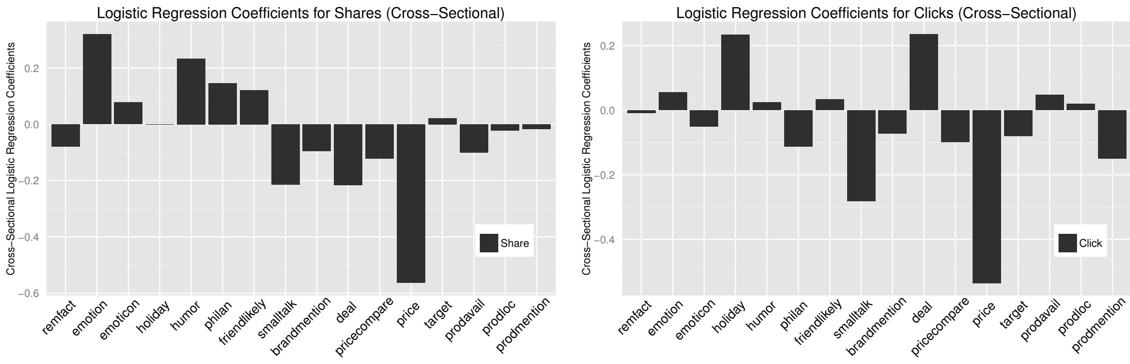

Figure 13: **Message Characteristic Coefficients for Shares and Click-throughs**: These bar graphs show the coefficients of logistic regression for EdgeRank-corrected models for Shares and Click-throughs. Only the significant coefficients are plotted.

4.4 Robustness and Sensitivity Analysis

Nonrandom Targeting of Posts by Firms to Consumers

We concentrated on EdgeRank-induced selection as the main difficulty in inference since we believe the specifics of the Facebook environment makes several other sources of confounds second-order compared to the effect of EdgeRank. One concern may be that firms may target content directly to specific users or sub-audiences on Facebook. In our context, such direct targeting is unlikely. In contrast to Facebook's banner advertisements or sponsored posts, the Facebook organic page environment (to which our data correspond) does not allow companies to target specific audiences. That is, any post by a firm is a candidate for all of its fans and Facebook determines which of these fans will see the post based on its proprietary algorithm. All targeting is implicitly implemented by Facebook via EdgeRank's filtering. The only platform factor that can be controlled by the firm is the time-of-day of release of the message.

Nevertheless, there are subtle ways in which targeting concerns may manifest themselves. One way is that firms observe that a particular type of content receives significant engagement, and subsequently start posting content that is similar to that. Thus, new content reflects past engagement generating an endogeneity concern related to behavioral targeting. Alternatively, certain kinds of firms may systematically pick particular kinds of content, so it may be unobserved firm effects that we are picking up in our estimates, and not content effects – another source of targeting-induced bias. To check whether our results reflect selection of particular types of posts by high-performing firms, we evaluate the diversity of content posted by firms as well as serial correlation in posts. To the extent that the diversity of posts by firms is limited, it raises the concern that certain kinds of content attributes were not used by some low or high performing firms. Similarly, if there is high serial correlation in posts by firms, it may reflect limited content diversity or that firms are choosing posts based on the performance of a previous post.

To evaluate this issue, we represent each post by a company by a binary vector of length sixteen (8 directly informative and 8 brand personality-related) in which 1 represents that the content attribute is present (and 0 otherwise). Next, all such vectors by the same firm are simply added up to form a vector indicating overall content creation for each firm. The Herfindahl index[13] is then calculated for each firm. The mean Herfindahl index is 0.089, and the median 0.088, which is just above the minimal possible value of $\frac{1}{16} = 0.0625$. Other concentration measures, such as Gini coefficient, also report similar patterns, which suggest that firms in our dataset post different types of content.

Similarly, to assess the extent of serial correlation in post content released by firms, we again represent each post with a vector of length sixteen, with each entry in the vector representing a binary variable for the presence of the 16 content attributes. For each firm, messages are ordered by release date, then the XOR function applied to all consecutive messages to measure content similarity.[14] A value of 1 for the XOR function means that a content attribute that was used in a previous post is not present in the current post, and vice versa. For example, if the first post released by a firm is the vector (1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0) and the second is (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1), then the XOR vector comparing the two is (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1). The mean of this XOR vector is $\frac{2}{16} = 0.125$. Computing the mean of such XOR vectors for all sequential pairs of posts for all firms in our data produced an average value of 0.32 (median 0.33). This suggests significant variation in content by firms in our dataset. Finally, firm-fixed effects are included in all specifications to guard against mixing up attribute driven differences in engagement with firm-specific effects.

A final issue is a concern that firms may be optimizing the timing of the messages. For example, say a firm wants to give an emotional message better exposure than a humorous message. If the firm knows that more people are logged on Facebook at 3pm versus 8am, the firm can post the emotional message at 3pm and the humorous one at 8am. Our model control for demographics and impressions to mitigate this issue on a first-order basis. We additionally check our data if there is evidence that firms are targeting certain times of the day to post specific content mixes. Our data show no evidence to support such timing choice. Figure 14 presents the distribution of the time of post for each of the sixteen content attributes. Each line represents a different content-type. This graph shows that while the volume of content has significant time-of-day dependence (such as at 5pm-6pm when people leave their workplace), the mix of content does not show such dependence. All distributions appear similar. In fact, none of the $\binom{16}{2}$ pair-wise Kolmogorov-Smirnov tests were able to reject the null that the lines came from the same distribution. The figure, coupled with high diversity of content, suggests that firms in our dataset were not systematically selecting content attributes by time-of-day. One may wonder what explains this apparent lack of sophistication. One reason may be due to the lack of social media analytics tools that provided content-level analytics to companies during the time of the data collection. In fact, social media analytics tools available in the market at the time of data collection only provided simple timing strategies such as what time of the day to post, as reflected in the

[13] A measure of diversity that ranges from $\frac{1}{n}$ to 1 where 1 means highly concentrated. For our case it can range from $\frac{1}{16}$ to 1 where $\frac{1}{16}$ means all sixteen contents are equally used.

[14] "Exclusive OR" – i.e., "A or B but not A and B".

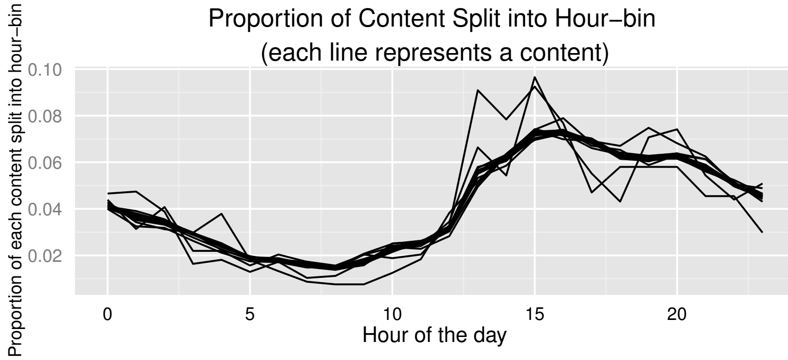

Figure 14: **Proportion of Content Posted Split into Hour-bin:** Each line represents one of sixteen content types.

data and controlled for in our model, but not which content attributes work well or when to post specific types of content. While it is hard to formally rule out a selection bias, these analyses of content diversity and timing suggests that firms' strategic targeting may of second order for our context.

Omitted Message Characteristics

A final concern is that engagement is driven by unmeasured message characteristics that co-occur with included message characteristics. To the extent that these unmeasured message characteristics drive engagement, they represents unobservables that are potentially correlated with included message characteristics and generate an omitted variables problem. This concern is plausible, but is second order in our view to the extent that we have included a very rich set of message characteristics. We have included and/or hard-coded a large number of message characteristics (including things like emotion and humor content, which are typically thought of as being in the unobservables). Our approach to this problem has been to convert unobservables into observables by collecting direct data on a relatively comprehensive set of message-characteristics and do direct inference on the effect of these hand-to-measure constructs. We assess the extent to which these omitted variables are problematic by using the residuals from the first-stage as an added control in the second-stage. To see this, note the residuals in Equation 3, $\epsilon_{kjt}^{(d)}$, represent unobserved reasons that users in demographic bucket d would be more likely to be targeted a message k by EdgeRank. As robustness, we ask whether our results on the effect of message attributes change when we control for these unobservable drivers of attractiveness of each bucket for that message. To do this, note that from our first-stage, we can obtain an estimate of the residual, denoted $\hat{\epsilon}_{kjt}^{(d)}$. We re-run our second stage estimation including the estimated $\hat{\epsilon}_{kjt}^{(d)}$-s as covariates in M_{kt} in Equation 5. We can interpret the revised results as the effect of message characteristics on engagement after "controlling for" the unobserved attractiveness of each bucket for that message. Results from these alternative models (presented in Appendices) show that the main qualitative features of our results are robust across these specifications.

Alternative Specifications

We also run a variety of alternative specifications to assess the robustness of our results. Estimates of these alternative specifications are presented in Appendices. First, we replicate the results using only the set of 5,000 messages directly coded up by the Amazon Mechanical Turkers. Second, we assess the extent to which the parameters are stable when we drop subsets of attributes. We find that the nature of our results remain unchanged.

5 Conclusions

We show through a large-scale study that content engineering in social media has a significant impact on user engagement as measured by *Likes*, comments, shares, and click-throughs for messages. Our analysis shows that brand personality-related content, such as emotional and philanthropic content, has a strong positive impact on engagement. This suggests that firms gain from sharing their brand personality and information about their social initiatives on social media. Further, we find that directly informative content has a negative impact on social media engagement, but certain informative content induce higher click-throughs. Thus, brand personality-related content primarily drives engagement and seems key for long term brand building, while directly informative content drives direct-response and seems key to performance marketing. This presents a challenge to marketers who seek to build a large following on social media and who seek to leverage that following to disseminate information about new products and promotions. One takeaway from our study is that these strategies work when directly informative content is combined with brand personality-related content to balance reach and engagement on the platform.

Because of the scale of our study (nearly 800 firms and more than 100,000 messages analyzed), we believe our results generalize and have broad applicability. Nonetheless, it is important to recognize several limitation of our study. First, we note that the results from any study on consumer response to content depend on the mix of content used in the study. For example, we find that messages mentioning holidays, especially by consumer product companies, have a negative effect on engagement. This may be due to excessive use of holiday messages by firms. It is possible that the effect may be positive if firms use these kinds of messages in moderation. Similarly, we find that emotional messages have a positive impact on engagement. Here again, it is possible this effect may reduce in the future if firms start using emotional content excessively. Hence, it is important to interpret our results in the context of the content mix used by firms and redo the analysis in the event of large-scale changes in the content mix used by firms. Ultimately, we urge managers to strike the right balance between the directly informative content (meant to drive leads and sales) and the brand personality-related content (meant to engage the consumers), especially since EdgeRank uses firms' current engagement level to determine future reach.

We used several metrics for user engagement, namely *Likes* and comments on messages as well as whether users share messages with friends or visit the link in the message. Our use of *Likes*, comments, shares, and click-throughs is motivated both by the widespread use of these metrics as marketing goals in social media settings, and also the availability of data. Future studies that evaluate other measures of interest can add value, particularly in validating the generalizability of our findings and in exploring mechanisms underpinning the effects we describe. As noted in the introduction, we do not address the question of how engagement affects product demand and firm's profits so as to complete the link between ad-attributes and those outcome measures. Such data are still not widely available at the scale needed for this study. Although it is not the focus of our study, it is worth highlighting that several extant studies have studied the link between Facebook engagements and sales, albeit at a smaller scale. For example, based on randomized studies, comScore (2012) reports a 38% lift in purchase for fans exposed to Starbucks on Facebook through Facebook Pages or Facebook paid advertising. Similarly, studies such as Kumar et al. (2013); Goh et al. (2013); Rishika et al. (2013); Li and Wu (2014); Miller and Tucker (2013); Sunghun et al. (2014); Luo and Zhang (2013); Luo et al. (2013) show that social media can be used to generate growth in sales, and ROI, consumer participation, retention, and profitability, connecting social media metrics such as "comments" to financial metrics.

The competition for consumer attention across media outlets is intense, especially on social media platforms. Consumers, in turn, are overwhelmed by the proliferation of online content, and it seems clear that marketers will not succeed without engineering this content for their audience. We hope this study contributes to improve content engineering by firms on social media sites and, more generally, creates interest in evaluating the effects of content on business outcomes.

References

90octane: 2012, 'New Facebook Content and ConversionStrategies Generate Engagement and Leads for Gates Automotive'. Technical report, 90octane.

Aaker, D. A.: 1996, *Building Strong Brands,*. New York: The Free Press.

Aaker, J.: 1997, 'Dimensions of brand personality'. *Journal of Marketing Research,* **34**(3), 347–356.

Abernethy, A. M. and G. R. Franke: 1996, 'The Information Content of Advertising: A Meta-Analysis'. *Journal of Advertising* **25**(2), 1–17.

Aggarwal, R., R. Gopal, R. Sankaranarayanan, and P. V. Singh: 2012, 'Blog, blogger, and the firm: Can negative employee posts lead to positive outcomes?'. *Information Systems Research* **23**(2), 306–322.

Ahn, H. and J. L. Powell: 1993, 'Semiparametric estimation of censored selection models with a nonparametric selection mechanism'. *Journal of Econometrics* **58**(3), 29.

Anand, B. and R. Shachar: 2009, 'Targeted advertising as a signal'. *Quantitative Marketing and Economics* **7**(3), 237–266.

Anand, B. and R. Shachar: 2011, 'Advertising the Matchmaker'. *Rand Journal of Economics* **42**(2), 205–245.

Anderson, S. and R. Renault: 2006, 'Advertising Content'. *American Economic Review* **March**, 93–113.

Armstrong, S.: 2010, *Persuasive Advertising.* Palgrave Macmillan.

Ascend2: 2013, 'Marketing Strategy Report: Social Media'. Technical report, Ascend2 and Research Underwriters.

Bagwell, K.: 2007, 'Chapter 28: The Economic Analysis of Advertising'. *Handbook of IO* **3**, 1701–1844.

Bennett, P. N.: 2006, 'Building Reliable Metaclassifiers for Text Learning'. Ph.D. thesis, CMU.

Berger, J.: 2011, 'Arousal Increases Social Transmission of Information'. *Psychological Science* **22**(7), 891–93.

Berger, J.: 2012, 'Word-of-Mouth: A Reveiew and Directions for Future Research'. *Work in Progress*.

Berger, J. and K. L. Milkman: 2012, 'What Makes online Content Viral?'. *Journal of Marketing Research* **49**(2), 192–205.

Berger, J. and E. Schwartz: 2011, 'What Drives Immediate and Ongoing Word-of-Mouth?'. *Journal of Marketing Research* **48**(5), 869–880.

Bertrand, M., D. Karlan, S. Mullianathan, E. Shafir, and J. Zinman: 2010, 'What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment'. pp. 263–306.

Bird, S., E. Klein, and E. Loper: 2009, *Natural Language Processing with Python*. O'Reilly Media.

Brieman, L.: 1996, 'Bagging predictors'. *Machine Learning* **24**, 123–140.

Buhrmester, M., T. Kwang, and S. D. Gosling: 2011, 'Amazon's Mechanical Turk: a New Source of Inexpensive, Yet High-Quality, Data?'. *Psychological Science* **6**(1).

Butters, G.: 1977, 'Equilibrium Distributions of Sales and Advertising Prices'. *Review of Economic Studies* **44**(3), 465–91.

Chadwick-Martin-Bailey: 2010, 'Consumers Engaged Via Social Media Are More Likely To Buy, Recommend'.

Chandy, R., G. Tellis, D. Macinnis, and P. Thaivanich: 2001, 'What to Say When: Advertising Appeals in Evolving Markets'. *Journal of Marketing Research* **38**, 399–414.

Cialdini, R.: 2001, *Influence: Science and Practice*. Needham Heights, MA: Allyn and Bacon.

comScore: 2012, 'The Power of Like 2'. Technical report, comScore.

comScore: 2013, 'The Power of Like Europe'. Technical report, comScore.

Creamer, M.: 2012, 'Study Only 1 percent of Facebook 'Fans' Engage With Brands'. *AdAge*.

Dellarocas, C.: 2006, 'Strategic manipulation of internet opinion forums: Implications for consumers and firms'. *Management Science* **52**(10), 1577–1593.

Ellickson, P. B. and S. Misra: 2010, 'Enriching Interactions: Incorporating Outcome Data into Static Discrete Games'. *Quantitative Marketing and Economics* **1**(10).

eMarketer: 2013, 'Which Social Media Marketing Tactics Work Best?'. Technical report, eMarketer.

Freund, Y. and R. E. Schapire: 1995, 'A decision-theoretic Generalization of on-line learning and an application to boosting'. *Computational Learning Theory: EUROCOLT* pp. 23–37.

Gardete, P. M.: 2013, 'Cheap-Talk Advertising and Misrepresentation in Vertically Differentiated Markets'. *Marketing Science*.

Gentzkow, M. and J. Shapiro: 2010, 'What Drives Media Slant? Evidence from U.S. Newspapers'. *Econometrica* **78**(1).

Gerber, S.: 2014, '9 Ways Social Media Marketing Will Change in 2014'.

Geva, T. and J. Zahavi: 2013, 'Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news'. *Decision Support Systems*.

Ghose, A., P. G. Ipeirotis, and B. Li: 2012, 'Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content'. *Marketing Science* **31**(3).

Goh, K.-Y., C.-S. Heng, and Z. Lin: 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.

Govers, P. C. and J. P. Schoormans: 2005, 'Product personality and its influence on consumer preference'. *Journal of Consumer Marketing* **22**(4), 189–197.

Grossman, G. and C. Shapiro: 1984, 'Informative Advertising with Differentiated Products'. *Review of Economic Studies* **51**(1), 63–81.

Hassan, M. H., B. S. Vikas, and H. Fiorletta: 2011, 'Accurate Information Extraction for Quantitaive Financial Events'. *CIKM*.

Hastie, T., R. Tibshirani, and J. Friedman: 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hastie, T. and R. J. Tibshirani: 1990, *Generalized Additive Models*. Chapman and Hall CRC.

Heckman, J. and R. Robb: 1986, *Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes*.

HubSpot: 2013, 'State of Inbound Marketing'. Technical report, HubSpot.

Ipeirotis, P. G., F. Provost, and J. Wang: 2010, 'Quality Mangagement on Amazon Mechanical Turk'. *HCOMP*.

Jabr, W., R. Mookerjee, Y. Tan, and V. S. Mookerjee: 2014, 'Leveraging Philanthropic Behavior for Customer Support: The Case of User Support Forums'. *MIS Quarterly* **38**(1).

Jurafsky, D. and J. H. Martin: 2008, *Speech and Language Processing*. Prentice Hall.

Keller, K. L.: 1993, 'Conceptualizing, measuring, and managing customer-based brand equity'. *The Journal of Marketing* **1**(1-22).

Kihlstrom, R. and M. Riordan: 1984, 'Advertising as a Signal'. *Journal of Political Economy* **92**(3), 427–50.

Kumar, V., V. Bhaskara, R. Mirchandani, and M. Shah: 2013, 'Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey'. *Marketing Science* **32**(2), 194–212.

Li, X. and L. Wu: 2014, 'Measuring Effects of Observational Learning and Social-Network Word-of-Mouth (WOM) on the Sales of Daily-Deal Vouchers'. *SSRN Working Paper*.

Liaukonyte, J., T. Teixeira, and K. Wilbur: 2013, 'TV Advertising Content and Online Shopping'. *(in progress)*.

Luo, X. and J. Zhang: 2013, 'How Do Consumer Buzz and Traffic in Social Media Marketing Predict the Value of the Firm?'. *Journal of Management Information Systems* **30**(2), 213–238.

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

Manchanda, P., P. E. Rossi, and P. Chintagunta.: 2004, 'Response Modeling with Non- Random Marketing Mix Variables'. *Journal of Marketing Research* **41**, 467–478.

Mason, W. A. and S. Suri: 2012, 'Conducting behavioral research on Amazon's Mechanical Turk'. *Behavior Research Methods* **44**(1), 1–23.

Mayzlin, D. and J. Shin: 2011, 'Uninformative Advertising as an Invitation to Search'. *Marketing Science* **30**(4), 666–685.

Milgrom, P. and J. Roberts: 1986, 'Price and Advertising Signals of Product Quality'. *Journal of Political Economy* **94**(4), 796–821.

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

Nair, H., S. Misra, W. Hornbuckle IV, R. Mishra, and A. Acharya: 2013, 'Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation'.

Nelson, P.: 1974, 'Advertising as Information'. *Journal of Political Economy* **82**(4), 729–54.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko: 2012, 'Mine Your Own Business: Market-Structure Surveillance Through Text Mining'. *Marketing Science* **31**(3), 521–543.

Oestreicher-Singer, G. and L. Zalmanson: 2013, 'CONTENT OR COMMUNITY? A DIGITAL BUSINESS STRATEGY FOR CONTENT PROVIDERS IN THE SOCIAL AGE.'. *MIS Quarterly* **37**(2).

Paolacci, G., J. Chandler, and P. G. Ipeirotis: 2010, 'Running Experiements on Amazon Mechcanical Turk'. *Judgement and Decision Making* **5**(5).

Porter, L. and G. J. Golan: 2006, 'From Subservient Chickens to Brawny Men: a Comparison of Viral Advertising to Television Advertising'. *Journal of Interactive Advertising* **6**(2).

Ragan and N. O. C. Solutions: 2012, 'Structuring a Social Media Team'. Technical report, Ragan and NASDAQ OMX Corporate Solutions.

Resnik, A. and B. L. Stern: 1977, 'An Analysis of Informative Content in Television Advertising'. *Journal of Marketing* **41**(1).

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

Rousseeuw, P. J.: 1987, 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis'. *Journal of computational and applied mathematics* **20**, 53–65.

Rui, H., Y. Liu, and A. Whinston: 2013, 'Whose and what chatter matters? The effect of tweets on movie sales'. *Decision Support Systems* **55**(4), 863–870.

Sahni, N., C. Wheeler, and P. Chintagunta: 2015, 'Advertising Tailoring: Persuasive effect of advertising content'. *Working paper, Stanford GSB.*

Sheng, V., F. Provost, and P. G. Ipeirotis: 2007, 'Get another label? Improving data quality and data mining using multiple, noisy labelers'. *14th ACM SIGKDD* pp. 614–622.

Singh, P. V., N. Sahoo, and T. Mukhopadhyay: 2014, 'Seeking variety: A dynamic model of employee blog reading behavior'. *forthcoming in Information Systems Research.*

SmartBrief: 2010, 'The State of Social Media for Business'. Technical report, SmartBrief.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng: 2008, 'Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks'. *Empirical Methods in Natural Language Processing.*

Sunghun, C., A. Animesh, G. Han, and A. Pinsonneault: 2014, 'Firms' Social Media Efforts, Consumer Behavior, and Firm Performance: Evidence from Facebook'. *SSRN Working Paper* **http://ssrn.com/abstract=2448510**.

Susarla, A., J.-H. Oh, and Y. Tan: 2012, 'Social networks and the diffusion of user-generated content: Evidence from YouTube'. *Information Systems Research* **23**(1), 23–41.

Tucker, C.: 2012a, 'Ad Virality and Ad Persuasiveness'. *MIT Working paper.*

Tucker, C.: 2012b, 'Social Advertising'. *Working paper, MIT.*

Vakratsas, D. and T. Ambler: 1999, 'How Advertising Works: What do we really know?'. *Journal of Marketing* **63**(1).

Weiss, M. and F. Huber: 2000, *The Value of Brand Personalities: The Phenomenon of the Strategic Positioning of Brands*. Wiesbaden: German University Publishing.

Wildfire: 2012, 'Edgerank and Your Brand: An Analysis of Messaging Engagement on Facebook by Business Category'. Technical report, Wildfire.

Wood, S. N.: 2006, *Generalized Additive Models: An Introduction with R*. CRC press.

WSJ: 2012, 'GM Says Facebook Ads Don't Pay Off'.

Xu, L. and A. Krzyzak: 1992, 'Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition'. *IEEE Transactions on systems*.

Zadrozny, B. and C. Elkan: 2002, 'Transforming Classifier Scores into Accurate Multiclass Probability Estimates'. *KDD*.

Appendix 1: Survey Instrument

We are interested in identifying the content of social media messages. Please read the following short message posted on a social media site and answer the following questions. Thank you.

{Message Content}

Indicate the message characteristics by (Yes/No).

The Message,

1: contains an interesting or remarkable fact.
2: includes emotional content.
3: uses humor.
4: includes **philanthropic, awareness or activist content** (promotes social welfare, social change, or informs the reader about certain facts to improve people's lives).
5: **mentions a company (organization) name or brand name** (e.g. Nike, Red Cross, VitaminWater or musicians such as The Beatles).
If yes, please type the company or brand name as it appears in the message in the following box (only one per box)
Most relevant company/brand mentioned ___________
Second company/brand mentioned (if any) ___________
6: provides information about any type of **discounts or freebies** (e.g., deals, coupons, promotional offers, sweepstakes, rewards, free items).
7: makes **a price comparison** or **a price-match guarantee** against competitors' product(s) or service(s).
8: contains content **other than** about a product or company business (e.g. small talk, social pleasantry, casual banter, "Happy Halloween", "what was your day like?").
9: **is targeted towards an audience segment.** E.g., a particular demographic (race, gender, age, location) or people with certain qualifications or characteristics. (e.g. "all moms, do you like Gerber?", "All chocolate lovers, get this deal").
10: provides information about the **availability of a product or service** (e.g. Only 5 deals remain, available only until tomorrow, new product coming tomorrow).
11: provides information about **where or how to obtain a product or service** (e.g. link, physical location to buy, general location)?
12: does the message have a dollar sign ($)?
13: Are your friends on social media likely to post messages such as the above?
14: The message is specific to a **particular product or service** (e.g. electronics, drinks, music, etc.).
If yes, please type the mentioned product or service:
Most relevant product mentioned ___________

Figure 15: **Survey Form Used in Amazon Mechanical Turk**

Appendix 2: Amazon Mechanical Turk - Robust Content Extraction

Following best-practices in the literature, we employ the following strategies to improve the quality of classification by the Turkers in our study.

1. For each message, at least 9 different Turkers' inputs are recorded. We obtain the final classification by a majority-voting rule.

2. We restrict the quality of Turkers included in our study to comprise only those with at least 100 reported completed tasks and 97% or better reported task-approval rates.

3. We use only Turkers from the US so as to filter out those potentially not proficient in English, and to closely match the user-base from our data (recall, our data has been filtered to only include pages located in the US).

4. We refined our survey instrument through an iterative series of about 10 pilot studies, in which we asked Turkers to identify confusing or unclear questions. In each iteration, we asked 10-30 Turkers to identify confusing questions and the reasons they found those questions confusing. We refined the survey in this manner till almost all queried Turkers stated no questions were confusing.

5. To filter out participants who were not paying attention, we included an easily verifiable test question "does the message have a dollar sign ($)?". Responses from Turkers that failed the verification test are dropped from the data.

6. In order to incentivize workers, we awarded additional bonuses of $2-$5 to the top 20 workers with exceptional accuracy and throughput.

7. On average, we found that message tagging took a little over 3 minutes and it typically took at least 20 seconds or more to completely read the tagging questions. We defined less than 30 seconds to be too short, and discarded any message tags with completion times shorter than that duration to filter out inattentive Turkers and automated programs ("bots").

8. Once a Turker tags more than 20 messages, a couple of tagged samples are randomly picked and manually examined for quality and performance. This process identified about 20 high-volume Turkers who completed all surveys in less than 10 seconds and tagged several thousands of messages (there were also Turkers who took time to complete the surveys but chose seemingly random answers). We concluded these were automated programs. These results were dropped, and the Turkers "hard blocked" from the survey, via the blocking option provided in AMT.

We believe our methodology for content-classification has strong external validity. The binary classification task that we serve to the AMT Turkers in our study is relatively simpler than the more complex tasks for which AMT-based data have been employed successfully in the literature. The existing AMT literature has documented evidence that several of the strategies implemented above improves the quality of the data generated (Mason and Suri (2012); Ipeirotis et al. (2010); Paolacci et al. (2010)). Snow et al. (2008) show that combining results from a few Turkers can produce data equivalent in quality to that of expert labelers for a variety of text tagging tasks. Similarly, Sheng et al. (2007) document that repeated labeling of the type we implement wherein each message is tagged by multiple Turkers, is preferable to single labeling in which one person tags one sentence. Finally, evaluating AMT based studies, Buhrmester et al. (2011) concludes that (1) Turkers are demographically more diverse than regular psychometric studies samples, and (2) the data obtained are at least as reliable as those obtained via traditional methods as measured by psychometric standards such as Cronbach's Alpha, a commonly used inter-rater reliability measure.

Appendix 3: NLP Algorithm

This section provides detailed outline of the algorithm used in the paper. Figure 16 shows the process visually.

Training The Algorithm

1. The raw textual data of $5,000$ messages in the training sample are broken down into basic building blocks of sentences using *stop-words removal* (removing punctuation and words with low information such as the definite article "the"), *tokenization* (the process of breaking a sentence into words, phrases, and symbols or "tokens"), *stemming* (the process of reducing inflected words to their root form, e.g., "playing" to "play"), and *part-of-speech tagging* (determining part-of-speech such as nouns). For reference see Jurafsky and Martin (2008). In this process, the input to the algorithm is a regular sentence and the output is an ordered set of fundamental linguistic entities with semantic values. We use a highly regarded python NLP framework named NLTK (Bird et al., 2009) to implement this step.

2. Once the messages are broken down as above, an algorithm extracts sentence-level attributes and sentence-structure rules that help identify the included content. Some examples of sentence-level attributes and rules include: frequent noun words (bag-of-words approach), bigrams, the ratio of part-of-speech used, tf-idf (term-frequency and inverse document frequency) weighted informative word weights, and whether "a specific key-word is present" rule. For completeness, we describe each of these in Table 8. The key to designing a successful NLP algorithm is to figure out what we (humans) do when identifying certain information. For example, what do we notice about the sentences we have identified as having emotional content? We may notice the use of certain types of words, use of exclamation marks, the use of capital letters, etc. At the end of this step, the dataset consists of sentence-level attributes generated as above (the x-variables), corresponding to a series of binary (content present/not-present) content labels generated from AMT (the y-variables).

3. For each binary content label, we then train a classification model by combining statistical and rule-based classifiers. In this step, the NLP algorithm fits the binary content label (the y-variable) using the sentence-level attributes as the x-variables. For example, the algorithm would fit whether or not a message has emotional content as tagged by AMT using the sentence attributes extracted from the message via step 2. We use a variety of different classifiers in this step including logistic regression with L1 regularization (which penalizes the number of attributes and is commonly used for attribute selection for problems with many attributes; see (Hastie et al., 2009)), Naive Bayes (a probabilistic classifier that applies Bayes theorem based on presence or absence of features), and *support vector machines* (a gold-standard algorithm in machine learning that works well for high dimensional problems) with different flavors of regularization and kernels [15]. To account for imbalance in positive and negative class

[15] We tried support vector machines with L1 and L2 regularization and various kernels including linear, radial basis function,

labels in some content, we utilized combination of class-weighted classifiers and resampling methods.

4. To train the ultimate predictive classifier, we use ensemble methods to combine results from the multiple statistical classifiers we fit in step 3. The motivation for ensemble learning is that different classifiers perform differently based on underlying characteristics of data or have varying precision or recall in different locations of the feature vector space. Thus, combining them will achieve better classification output either by reducing variance (e.g. Bagging (Brieman, 1996)) or reducing bias (e.g. Boosting (Freund and Schapire, 1995)). Please see Xu and Krzyzak (1992); Bennett (2006) for further reading on ensemble methods. This step involves combining the prediction from individual classifiers by weighted-majority voting, unweighted-majority voting, or a more elaborate method called isotonic regression (Zadrozny and Elkan, 2002) and choosing the best performing method in terms of accuracy, precision and recall for each content profiles. In our case, we found that support vector machine based classifiers delivered high precision and low recall, while Naive Bayes based classifiers delivered high recall but low precision. By combining these, we were able to develop an improved classifier that delivers higher precision and recall and in effect, higher accuracy. Table 9 shows the improvement of the final ensemble learning method relative to using only one support vector machine. As shown, the gains from combining classifiers are substantial. We obtain similar results for negative class labels.

5. Finally, we assess the performance of the overall NLP algorithm on three measures, viz., accuracy, precision, and recall (as defined in Footnote 4) using the "10-fold cross-validation" method. Under this strategy, we split the data randomly into 10 equal subsets before the step 2. One of the subsets is used as the validation sample, and the algorithm trained on the remaining 9 sets. This is repeated 10 times, each time using a different subset as the validation sample, and the performance measures averaged across the 10 runs. The use of 10-fold cross-validation reduces the risk of overfitting and increases the external validity of the NLP algorithm we develop. Note, 10-fold cross-validation of this sort is computationally intensive and impacts performance measures negatively and is not implemented in some existing papers in business research. While the use of 10-fold cross-validation may negatively impact the performance measures, it is necessary to increase external validity. Table 9 shows these metrics for different content profiles. The performance is extremely good and comparable to performance achieved by the leading financial information text mining systems (Hassan et al., 2011).

6. We repeat steps 2-5 until desired performance measures are achieved.

Tagging New Messages

1. For each new messages repeat steps 1-2 described above.

2. Use the ultimate classifier developed above to predict whether a particular type of content is present or not.

and polynomial kernels. For more details, refer to Hastie et al. (2009).

One can think of this NLP algorithm as emulating the Turkers' collective opinion in content-coding.

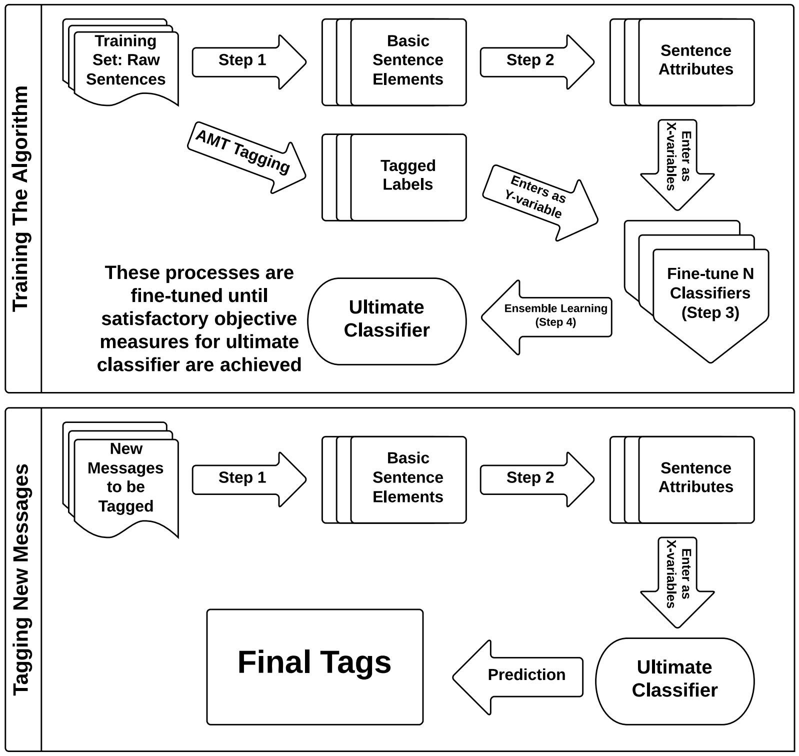

Figure 16: **Diagram of NLP Training and Tagging Procedure:** This diagram shows the steps of training the NLP algorithm and using the algorithm to tag the remaining messages. These steps are described in Appendix 3.

Rules and Attributes	**Description**
Bag of Words	Collects all the words and frequency for a message. Different variations include collecting top N most occurring words.
Bigram	A bigram is formed by two adjacent words (e.g. "Bigram is", "is formed" are bigrams).
Ratio of part-of-speech	Part-of-speech (noun, verb, etc) ratio in each message.
TF-IDF weighted informative word	Term-Frequency and Inverse Document Frequency weighs each word based on their occurrence in the entire data and in a single message.
Specific Keywords	Specific keywords for different content can be collected and searched. e.g., Philanthropic messages have high change of containing the words "donate" and "help". For brand and product identification, large online lists were scraped and converted into dictionaries for checking.
Frequency of different punctuation marks	Counts the number of different punctuations such as exclamation mark and question mark. This helps to identify emotion, questions, appearance of deals etc.
Count of non-alphanumerics	Counts the number of characters that are not A-Z and 0-9.

Table 8: **A Few Examples of Message Attributes Used in Natural Language Processing Algorithm**

	With Ensemble Learning (The Best Performing Algorithm)				Without Ensemble Learning (Support Vector Machine version 1 + Rule-based)		
	Accuracy	Precision	Recall		Accuracy	Precision	Recall
REMFACT	0.94	0.99	0.68		0.88	0.99	0.33
EMOTION	0.97	0.99	0.87		0.94	0.98	0.65
HUMOR	0.98	1	0.90		0.97	1	0.14
PHILANTHROPIC	0.97	0.99	0.85		0.93	0.99	0.62
FRIENDLIKELY	0.94	0.99	0.68		0.90	0.99	0.41
SMALLTALK	0.85	0.88	0.80		0.78	0.34	0.28
DEAL	0.94	0.99	0.65		0.90	1	0.43
PRICECOMPARE	0.99	0.99	1		0.99	1	0.85
TARGETING	0.98	0.99	0.89		0.95	0.99	0.71
PRODAVAILABILITY	0.96	0.99	0.76		0.91	1	0.10
PRODLOCATION	0.97	0.99	0.90		0.87	1	0.11

Table 9: **Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation**: This table presents metrics for performance of the classification algorithms used. The left 3 columns show the metrics for the final algorithm which combines classifiers via ensemble learning method while the right 3 columns show the metric for a support vector machine algorithm. Notice that the support vector machine classifier tends to have low recall and high precision. Naive Bayes tends to have high recall but low precision. Classifiers on their own cannot successfully overcome precision-recall tradeoff (if one is higher, one is lower). But combining many different classifiers with ensemble learning can increase both precision and recall. We obtain similar results for negative class labels.

Appendix 4: Obtaining the Number of Impressions

We discuss our procedure for constructing $n_{kjt}^{(d)}$, the number of impressions for each message k of firm j in day t split by demographic bin d for use in the EdgeRank correction model. As mentioned above, $n_{kjt}^{(d)}$ is not directly reported by Company X (or made available to page-owners by Facebook *Insights*). Instead, Company X reports $n_{jt}^{(d)}$, the number of impressions for all messages associated with firm j in demographic bucket d on day t, which is essentially $n_{kjt}^{(d)}$ summed across all k associated with j. In addition, we observe n_{kjt}, the total number of impression obtained by a specific post k by firm j on day t. To assess how we may split this across the various demographic buckets, we checked the extent to which pages release different types of messages over time. The bulk of impressions for a message occur within the first week of its release. Hence, the total impressions for a page on a given day out of a specific demographic bucket, $n_{jt}^{(d)}$, reflects the aggregate impressions to users in that bucket of all messages released by that firm over the past one week. Since EdgeRank allocates messages to users by message-type, if the firm releases the same *type of messages* (i.e., photos, videos, status updates, apps or links) over a week's duration, then the the distribution of $n_{jt}^{(d)}$ across the various k messages released by firm j within the past week may be roughly the same. In other words, the distribution of demographics of the impressions of all messages released by a firm in the past week should be the same if all those released messages are similar.

To check informally if this is the case, we picked a random sample of $10,000$ page-7-day combinations from our data. For each combination, we collated all the messages released by that page during that 7-day window and tabulated the type of these messages (i.e., photos, videos, status updates, apps or links). We then construct two concentration metrics, $\mathbb{C}_1$, the proportion of messages released by that page during that 7-day window that belong to the highest type bucket, and $\mathbb{C}_2$, the proportion of messages released by that page during that 7-day window that belong to the highest and second-highest type bucket. $\mathbb{C}_1$ and $\mathbb{C}_2$ are analogous to top-firm and top-two-firm concentration ratios used in industry-concentration studies, and measure the extent to which the messages released by a page in a given 7-day period are spread across types. If all messages released by a page during that 7-day window are of the same type, $\mathbb{C}_1$ and $\mathbb{C}_2$ will be both 1. The spread away from 1 thus indicates higher variation in message-types released by an average Facebook page over a week's duration. Table 10 reports on the distribution of $\mathbb{C}_1$ and $\mathbb{C}_2$ we computed in this manner. Looking at Table 10, we find that the median $\mathbb{C}_1$ is .71 (mean .72) and the median $\mathbb{C}_2$ is 1.0 (mean .94). Most pages seem to be releasing at-most 2-types of messages within a week window, and more than $2/3^{rd}$ of messages released by an average page in an average week are of the same type. Given this, we assume that n_{kjt} is split into $n_{kjt}^{(d)}$ with the same distribution as given by $n_{jt}^{(d)}$. We construct the variable $n_{kjt}^{(d)}$ in the left hand-side of the EdgeRank correction equation 2 in this manner. We replicated the study with different split distributions such as 7-day aggregated distribution of $n_{jt}^{(d)}$, kernel smoothed distribution, and the results were similar.

The method is not without its limitations. We view it as a practical way to deal with the lack of data-reporting by Facebook, while exploiting the variation embedded in the observed impressions and to correlate

	Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
$\mathbb{C}_1$	0.250	0.535	0.706	0.719	0.915	1.000
$\mathbb{C}_2$	0.500	0.889	1.000	0.942	1.000	1.000

Table 10: **Distribution of the Top ($\mathbb{C}_1$) and Top-two ($\mathbb{C}_2$) Concentration Ratios of the type of messages Served by a Facebook Page over a Randomly picked 7-day period**

it with the observed variation in the the potential market for each message in each demographic bucket $\left(\mathbb{N}_{jt}^{(d)}\right)$. The method produces potential measurement error in the dependent variable, $n_{kjt}^{(d)}$ in the EdgeRank correction stage. Measurement error in the dependent variable is absorbed into the RHS unobservables and is usually less of a concern unless it is correlated with the unobservables. The fact that we include page-fixed effects separately for each demographic ($\theta_{1j}^{(d)}$ in Equation 3) mitigates concerns to a large extent that these unobservables may systematically be correlated with included characteristics.

Appendix 5: Results Across Different Industries

In order to investigate how industry moderate the impact of content on engagement, we repeated the main analysis on subsetted data on each industry type as described and categorized in Table 2.

Figure 17 shows the results on content effects by industry. Only the statistically significant results are graphed and all results are EdgeRank-corrected. Some coefficients are different across industries both in magnitude and, for some variables, in direction. For example, emotional and philanthropic content has the most positive impact on Facebook pages of type “Organizations” which include non-profits, educational organizations and religious groups. Further, while mentioning holidays has a negative impact on engagement for most industry types, it has a positive impact on engagement for Organizations. Similarly, looking at informative content, we observe that variables such as Price, Product Availability, and Product Mentions generally have a negative impact on engagement for most industry types, but have a positive impact for industry type “Celebrity.” Users seem more forgiving of celebrity pages endorsing products and sharing price information. On the other hand, there are content that have persistent effects across the industry. For example, post type, photo, has positive engagement across the industry, while including blanks have positive impact on comment for all industries.

Comparing Figures 3 and 17 also provides interesting comparisons of what each industry is currently posting and what users engage with. For example, pages of types Places and Businesses, Entertainment, and Consumer Products do not post emotional content much though Figure 17 shows that emotional content induce higher *Likes* and Comments. Similarly, while Places and Business pages tend to post more of deal content, only Consumer Product pages seem to be benefiting from the deal content (in terms of obtaining more comments). Places and Businesses pages also post larger percent of product availability content while only the Consumer Product and Celebrity pages benefit from inclusion of such content.

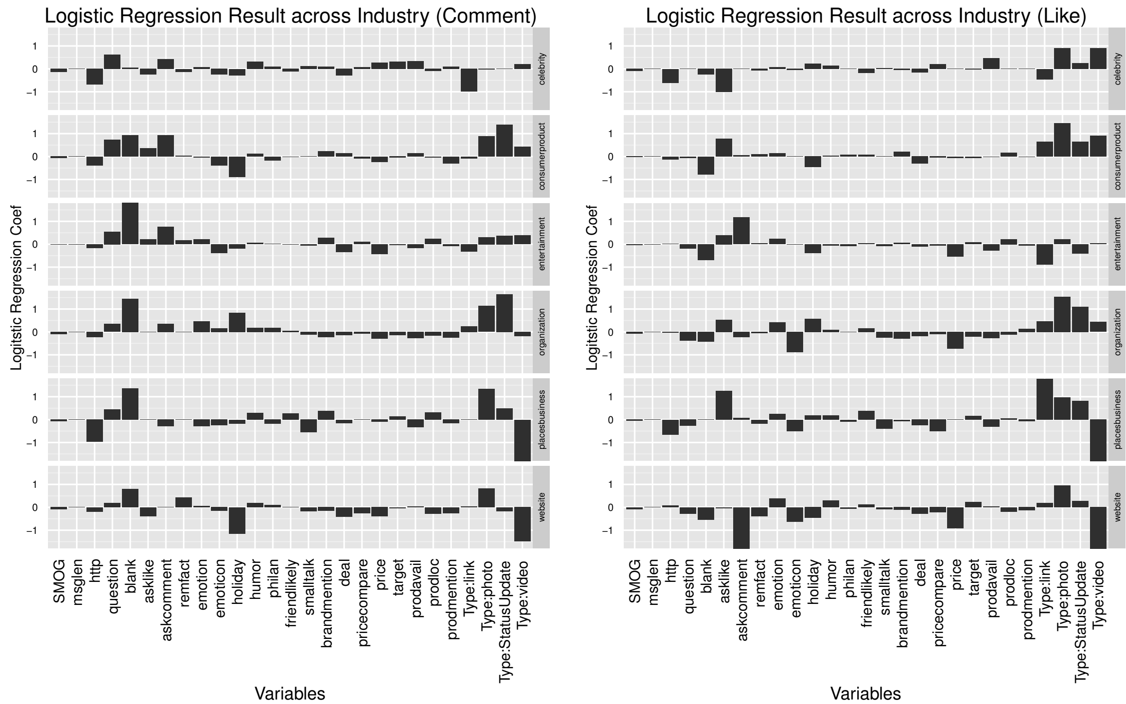

Figure 17: **Logistic Regression by Industry (Comments and *Likes*):** This bar graphs show the coefficients of logistic regression for EdgeRank-corrected model. Only the significant coefficients ($p<0.05$) are graphed. In the *Like* plot on the right, the coefficient for ASKCOMMENT for websites is -4.8 but zoomed in to optimize the clarity of the graph.

Appendix 6: Result for AMT-Tagged & Different Models

	NO ER COMMENT	OR	ER COMMENT	OR	NO ER LIKE	OR	ER LIKE	OR
Constant	-5.431*** (0.047)		-5.682*** (0.041)		-3.326*** (0.014)		-4.199*** (0.012)	
SMOG	-0.098*** (0.001)	0.907	-0.125*** (0.001)	0.882	-0.045*** (0.000)	0.956	-0.066*** (0.000)	0.936
MSGLEN	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000
HTTP	-0.253*** (0.005)	0.776	-0.143*** (0.005)	0.867	-0.097*** (0.002)	0.908	-0.025*** (0.002)	0.975
QUESTION	0.164*** (0.004)	1.178	0.191*** (0.004)	1.210	-0.325*** (0.001)	0.723	-0.220*** (0.001)	0.803
BLANK	0.739*** (0.012)	2.094	0.755*** (0.012)	2.128	-1.009*** (0.010)	0.365	-1.098*** (0.010)	0.334
ASKLIKE	-0.376*** (0.033)	0.687	-0.336*** (0.033)	0.715	-0.100*** (0.009)	0.905	-0.072*** (0.009)	0.931
ASKCOMMENT	0.927*** (0.049)	2.527	0.272*** (0.049)	1.313	0.111*** (0.020)	1.117	-0.406*** (0.020)	0.666
Brand personality-related								
REMFACT	0.105*** (0.006)	1.111	0.145*** (0.005)	1.156	-0.063*** (0.002)	0.939	-0.032*** (0.002)	0.969
EMOTION	0.159*** (0.006)	1.172	0.135*** (0.006)	1.145	0.084*** (0.002)	1.088	0.106*** (0.002)	1.112
EMOTICON	-0.383*** (0.026)	0.682	-0.361*** (0.026)	0.697	-0.416*** (0.008)	0.660	-0.192*** (0.008)	0.825
HOLIDAYMENTION	-1.465*** (0.033)	0.231	-1.293*** (0.033)	0.274	-0.433*** (0.006)	0.649	-0.266*** (0.006)	0.766
HUMOR	-0.086*** (0.012)	0.918	0.163*** (0.012)	1.177	-0.275*** (0.004)	0.760	0.019*** (0.004)	1.019
PHILANTHROPIC	0.142*** (0.007)	1.153	0.138*** (0.007)	1.148	0.182*** (0.002)	1.200	0.168*** (0.002)	1.183
FRIENDLIKELY	-0.159*** (0.005)	0.853	-0.130*** (0.005)	0.878	0.206*** (0.002)	1.229	0.159*** (0.002)	1.172
SMALLTALK	-0.005 (0.004)	0.995	-0.051*** (0.004)	0.950	0.022*** (0.001)	1.022	-0.028*** (0.001)	0.972
Directly informative								
BRANDMENTION	-0.179*** (0.004)	0.836	-0.228*** (0.004)	0.796	-0.141*** (0.001)	0.868	-0.224*** (0.001)	0.799
DEAL	-0.151*** (0.008)	0.860	0.011 (0.007)	1.011	-0.478*** (0.002)	0.620	-0.302*** (0.002)	0.739
PRICECOMPARE	0.688*** (0.193)	1.990	0.302 (0.193)	1.353	-0.498*** (0.103)	0.608	-0.857*** (0.103)	0.424
PRICE	-0.051*** (0.014)	0.950	-0.500*** (0.014)	0.607	-0.316*** (0.005)	0.729	-0.712*** (0.005)	0.491
TARGET	0.188*** (0.010)	1.207	0.076*** (0.010)	1.079	0.130*** (0.003)	1.139	0.059*** (0.003)	1.061
PRODAVAIL	-0.281*** (0.007)	0.755	-0.360*** (0.007)	0.698	0.073*** (0.002)	1.076	-0.022*** (0.002)	0.978
PRODLOCATION	-0.151*** (0.006)	0.860	-0.015** (0.005)	0.985	-0.239*** (0.002)	0.787	-0.080*** (0.002)	0.923
PRODMENTION	-0.170*** (0.005)	0.844	-0.349*** (0.005)	0.705	0.291*** (0.001)	1.338	0.095*** (0.001)	1.100
Message Type - App is the base								
-Link	0.226*** (0.008)	1.254	-0.051*** (0.008)	0.950	0.004 (0.003)	1.004	-0.310*** (0.003)	0.733
-Photo	0.633*** (0.008)	1.883	0.379*** (0.008)	1.461	0.718*** (0.002)	2.050	0.418*** (0.002)	1.519
-Status Update	1.325*** (0.009)	3.762	1.148*** (0.009)	3.152	0.671*** (0.003)	1.956	0.553*** (0.003)	1.738
-Video	-0.191*** (0.027)	0.826	-0.447*** (0.027)	0.640	-0.636*** (0.011)	0.529	-0.928*** (0.011)	0.395
Industry Category - Celebrity is the base								
-ConsumerProduct	-0.024*** (0.007)	0.976	-0.368*** (0.007)	0.692	-0.541*** (0.002)	0.582	-0.761*** (0.002)	0.467
-Entertainment	0.014* (0.006)	1.014	0.221*** (0.006)	1.247	-0.472*** (0.002)	0.624	-0.169*** (0.002)	0.845
-Organization	0.222*** (0.007)	1.249	0.072*** (0.007)	1.075	-0.163*** (0.002)	0.850	-0.161*** (0.002)	0.851
-PlaceBusiness	0.600*** (0.014)	1.822	0.339*** (0.014)	1.404	-0.698*** (0.007)	0.498	-0.869*** (0.007)	0.419
-Websites	-0.076*** (0.009)	0.927	0.152*** (0.008)	1.164	0.099*** (0.002)	1.104	0.398*** (0.002)	1.489
McFadden R-sq.	0.288		0.239		0.295		0.214	
Nagelkerke R-sq.	0.288		0.239		0.296		0.216	
Log-likelihood	-299060.434		-428096.493		-1728807.768		-3,119,513.948	
Deviance	552293.46		810390.763		3365047.493		6,146,513.932	
AIC	598244.869		856316.985		3457739.535		6,239,151.895	
N	38706		38706		38706		38706	
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1								

Table 11: **Aggregate Logistic Regression Results For Comments and *Likes* (5000 Messages):** This table presents the aggregate logistic regression on comments and *Likes* for both EdgeRank-corrected (ER) and uncorrected (NO ER) for 5000 messages data tagged by Turkers. OR means Odds ratio and shows the odds ratio for the estimates left of the column.

	NO ER COMMENT	OR	ER COMMENT	OR	NO ER LIKE	OR	ER LIKE	OR
Constant	-8.122*** (0.004)		-7.129*** (0.004)		-5.168*** (0.001)		-4.205*** (0.001)	
SMOG	-0.047*** (0.000)	0.954	-0.059*** (0.000)	0.943	-0.031*** (0.000)	0.969	-0.049*** (0.000)	0.952
MSGLEN	0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000
HTTP	-0.545*** (0.002)	0.580	-0.301*** (0.002)	0.740	-0.388*** (0.000)	0.678	-0.085*** (0.000)	0.919
QUESTION	0.488*** (0.001)	1.629	0.581*** (0.001)	1.788	-0.284*** (0.000)	0.753	-0.186*** (0.000)	0.830
BLANK	1.002*** (0.003)	2.724	1.186*** (0.003)	3.274	-0.709*** (0.002)	0.492	-0.538*** (0.002)	0.584
ASKLIKE	0.080*** (0.010)	1.083	0.259*** (0.010)	1.296	0.494*** (0.003)	1.639	0.598*** (0.003)	1.818
ASKCOMMENT	0.514*** (0.021)	1.672	0.413*** (0.021)	1.511	-0.244*** (0.011)	0.783	-0.557*** (0.011)	0.573
Brand personality-related								
REMFACT	-0.026*** (0.002)	0.974	0.017*** (0.002)	1.017	-0.055*** (0.001)	0.946	-0.015*** (0.001)	0.985
EMOTION	0.213*** (0.002)	1.237	0.217*** (0.002)	1.242	0.218*** (0.001)	1.244	0.191*** (0.001)	1.210
EMOTICON	0.171*** (0.004)	1.186	0.101*** (0.004)	1.106	-0.056*** (0.001)	0.946	-0.087*** (0.001)	0.917
HOLIDAYMENTION	-0.485*** (0.014)	0.616	-0.378*** (0.014)	0.685	-0.329*** (0.004)	0.720	-0.162*** (0.004)	0.850
HUMOR	0.030*** (0.001)	1.030	0.094*** (0.002)	1.099	-0.049*** (0.000)	0.952	0.047*** (0.000)	1.048
PHILANTHROPIC	0.182*** (0.002)	1.200	0.261*** (0.002)	1.298	0.019*** (0.001)	1.019	0.157*** (0.001)	1.170
FRIENDLIKELY	0.008*** (0.002)	1.008	-0.053*** (0.002)	0.948	0.076*** (0.001)	1.079	0.005*** (0.001)	1.005
SMALLTALK	0.059*** (0.002)	1.061	-0.111*** (0.002)	0.895	-0.061*** (0.001)	0.941	-0.172*** (0.001)	0.842
Directly informative								
BRANDMENTION	0.013*** (0.002)	1.013	0.084*** (0.002)	1.088	-0.015*** (0.000)	0.985	0.043*** (0.000)	1.044
DEAL	-0.134*** (0.002)	0.875	-0.245*** (0.002)	0.783	-0.183*** (0.001)	0.833	-0.357*** (0.001)	0.700
PRICECOMPARE	-0.038*** (0.001)	0.963	0.008*** (0.001)	1.008	-0.033*** (0.000)	0.968	-0.020*** (0.000)	0.980
PRICE	-0.031*** (0.005)	0.969	-0.303*** (0.005)	0.739	-0.185*** (0.001)	0.831	-0.459*** (0.001)	0.632
TARGET	-0.021*** (0.002)	0.979	-0.135*** (0.002)	0.874	0.036*** (0.001)	1.037	-0.120*** (0.001)	0.887
PRODAVAIL	-0.090*** (0.002)	0.914	-0.003 (0.002)	0.997	-0.119*** (0.001)	0.888	0.038*** (0.001)	1.039
PRODLOCATION	-0.056*** (0.002)	0.946	-0.043*** (0.002)	0.958	0.072*** (0.001)	1.075	0.034*** (0.001)	1.035
PRODMENTION	-0.087*** (0.002)	0.917	-0.128*** (0.002)	0.880	0.076*** (0.001)	1.079	0.039*** (0.001)	1.040
Message Type - App is the base								
-Link	0.201*** (0.003)	1.223	-0.301*** (0.003)	0.740	0.116*** (0.001)	1.123	-0.389*** (0.001)	0.678
-Photo	0.876*** (0.003)	2.401	0.318*** (0.003)	1.374	1.036*** (0.001)	2.818	0.475*** (0.001)	1.608
-Status Update	1.118*** (0.003)	3.059	0.666*** (0.003)	1.946	0.458*** (0.001)	1.581	-0.015*** (0.001)	0.985
-Video	0.641*** (0.009)	1.898	0.108*** (0.009)	1.114	0.218*** (0.003)	1.244	0.203*** (0.003)	1.225
Industry Category - Celebrity is the base								
-ConsumerProduct	0.068*** (0.002)	1.070	-0.062*** (0.002)	0.940	-0.446*** (0.001)	0.640	-0.544*** (0.001)	0.580
-Entertainment	0.420*** (0.002)	1.522	0.660*** (0.002)	1.935	-0.260*** (0.001)	0.771	-0.067*** (0.001)	0.935
-Organization	0.497*** (0.002)	1.644	0.631*** (0.002)	1.879	-0.005*** (0.001)	0.995	0.110*** (0.001)	1.116
-PlaceBusiness	0.381*** (0.005)	1.464	0.234*** (0.005)	1.264	-0.726*** (0.002)	0.484	-0.933*** (0.002)	0.393
-Websites	0.218*** (0.003)	1.244	0.380*** (0.003)	1.462	0.183*** (0.001)	1.201	0.314*** (0.001)	1.369
1st Stage Residual	-0.000*** (0.000)	1.000	0.000*** (0.000)	1.000	-0.000*** (0.000)	1.000	0.000*** (0.000)	0.000
McFadden R-sq.	0.229		0.199		0.298		0.270	
Nagelkerke R-sq.	0.230		0.199		0.298		0.272	
Log-likelihood	-2575187.42		-3460249.739		-14563950.18		-24911467.22	
Deviance	4745694.741		6516530.605		28179842.5		48875491.61	
AIC	5150474.839		6920599.479		29128000.36		49823034.45	
N	665916		665916		665916		665916	
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1								

Table 12: **Aggregate Logistic Regression Results For Comments and *Likes* (All Messages with the 1st Stage Residuals as a Control Function):** This table presents the aggregate logistic regression on comments and *Likes* for both EdgeRank-corrected (ER) and uncorrected (NO ER) for all messages data. OR means Odds ratio and shows the odds ratio for the estimates left of the column. This second-stage model includes first-stage residuals as a control function.

Variable	Intercept only	Controls	Friendlikely	Brand personality-related	Directly informative	All (No demo controls)
Constant	-7.052***(0.001)	-6.906***(0.003)	-6.952***(0.004)	-6.975***(0.004)	-6.861***(0.004)	-6.889***(0.004)
SMOG		-0.065***(0.000)	-0.065***(0.000)	-0.065***(0.000)	-0.067***(0.000)	-0.068***(0.000)
MSGLEN		-0.000***(0.000)	-0.000***(0.000)	-0.000***(0.000)	-0.000***(0.000)	-0.000***(0.000)
HTTP		-0.406***(0.001)	-0.400***(0.001)	-0.393***(0.001)	-0.390***(0.002)	-0.355***(0.002)
QUESTION		0.543***(0.001)	0.546***(0.001)	0.557***(0.001)	0.541***(0.001)	0.564***(0.001)
BLANK		1.172***(0.003)	1.168***(0.003)	1.187***(0.003)	1.144***(0.003)	1.146***(0.003)
ASKLIKE		0.222***(0.010)	0.229***(0.010)	0.220***(0.010)	0.223***(0.010)	0.221***(0.010)
ASKCOMMENT		0.508***(0.021)	0.529***(0.021)	0.465***(0.021)	0.506***(0.021)	0.469***(0.021)
Message Type - App is the base						
-Link		-0.388***(0.003)	-0.389***(0.003)	-0.394***(0.003)	-0.376***(0.003)	-0.370***(0.003)
-Photo		0.376***(0.003)	0.371***(0.003)	0.365***(0.003)	0.372***(0.003)	0.373***(0.003)
-Status Update		0.656***(0.003)	0.654***(0.003)	0.644***(0.003)	0.645***(0.003)	0.641***(0.003)
-Video		0.406***(0.009)	0.406***(0.009)	0.410***(0.009)	0.395***(0.009)	0.398***(0.009)
Industry Category - Celebrity is the base						
-ConsumerProduct		-0.402***(0.002)	-0.392***(0.002)	-0.371***(0.002)	-0.374***(0.002)	-0.347***(0.002)
-Entertainment		0.514***(0.002)	0.519***(0.002)	0.537***(0.002)	0.514***(0.002)	0.529***(0.002)
-Organization		0.391***(0.002)	0.393***(0.002)	0.409***(0.002)	0.400***(0.002)	0.408***(0.002)
-PlaceBusiness		-0.087***(0.005)	-0.082***(0.005)	-0.052***(0.005)	-0.065***(0.005)	-0.021***(0.005)
-Websites		0.123***(0.002)	0.134***(0.002)	0.150***(0.003)	0.136***(0.003)	0.182***(0.003)
FRIENDLIKELY			0.064***(0.001)	-0.051***(0.002)		-0.006***(0.002)
REMFACT				-0.065***(0.002)		0.014***(0.002)
EMOTION				0.217***(0.002)		0.256***(0.002)
EMOTICON				0.109***(0.004)		0.121***(0.004)
HOLIDAYMENTION				-0.391***(0.014)		-0.388***(0.014)
HUMOR				0.022***(0.001)		0.072***(0.002)
PHILANTHROPIC				0.054***(0.002)		0.174***(0.002)
SMALLTALK				-0.055***(0.002)		-0.086***(0.002)
BRANDMENTION					0.071***(0.002)	0.081***(0.002)
DEAL					-0.124***(0.002)	-0.172***(0.002)
PRICECOMPARE					0.043***(0.001)	-0.006***(0.001)
PRICE					-0.368***(0.005)	-0.317***(0.005)
TARGET					0.049***(0.002)	-0.071***(0.002)
PRODAVAIL					0.028***(0.002)	-0.064***(0.002)
PRODLOCATION					0.023***(0.002)	0.011***(0.002)
PRODMENTION					-0.186***(0.002)	-0.151***(0.002)
						Demographic proportion variables excluded
McFadden R-sq.		0.161	0.161	0.165	0.164	0.171
Nagelkerke R-sq.		0.161	0.162	0.165	0.165	0.172
Log-likelihood	-4267283.759	-3612205.117	-3610817.72	-3597891.814	-3598824.088	-3570184.37
Deviance	8130598.643	6820441.36	6817666.566	6791814.753	6793679.301	6736399.867
AIC	8534569.517	7224474.234	7221701.44	7195863.627	7197730.175	7140466.74
N	665916	665916	665916	665916	665916	665916
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1						

Table 13: **Logistic Regression EdgeRank-Corrected Estimates Model Comparison (Comments)**

Variable	Intercept only	Controls	Friendlikely	Brand personality-related	Directly informative	All (No demo controls)
Constant	-4.662***(0.000)	-3.929***(0.001)	-3.979***(0.001)	-3.931***(0.001)	-3.921***(0.001)	-3.892***(0.001)
SMOG		-0.057***(0.000)	-0.057***(0.000)	-0.057***(0.000)	-0.060***(0.000)	-0.061***(0.000)
MSGLEN		-0.001***(0.000)	-0.000***(0.000)	-0.001***(0.000)	-0.000***(0.000)	-0.000***(0.000)
HTTP		-0.214***(0.000)	-0.207***(0.000)	-0.205***(0.000)	-0.205***(0.000)	-0.189***(0.000)
QUESTION		-0.198***(0.000)	-0.195***(0.000)	-0.181***(0.000)	-0.198***(0.000)	-0.175***(0.000)
BLANK		-0.603***(0.002)	-0.612***(0.002)	-0.592***(0.002)	-0.602***(0.002)	-0.596***(0.002)
ASKLIKE		0.576***(0.003)	0.583***(0.003)	0.576***(0.003)	0.564***(0.003)	0.565***(0.003)
ASKCOMMENT		-0.379***(0.011)	-0.359***(0.011)	-0.444***(0.011)	-0.404***(0.011)	-0.465***(0.011)
Message Type - App is the base						
-Link		-0.517***(0.001)	-0.519***(0.001)	-0.518***(0.001)	-0.511***(0.001)	-0.505***(0.001)
-Photo		0.572***(0.001)	0.567***(0.001)	0.561***(0.001)	0.566***(0.001)	0.561***(0.001)
-Status Update		-0.062***(0.001)	-0.067***(0.001)	-0.083***(0.001)	-0.067***(0.001)	-0.083***(0.001)
-Video		0.244***(0.003)	0.245***(0.003)	0.252***(0.003)	0.225***(0.003)	0.231***(0.003)
Industry Category - Celebrity is the base						
-ConsumerProduct		-0.948***(0.001)	-0.937***(0.001)	-0.941***(0.001)	-0.939***(0.001)	-0.932***(0.001)
-Entertainment		-0.172***(0.001)	-0.168***(0.001)	-0.174***(0.001)	-0.189***(0.001)	-0.193***(0.001)
-Organization		-0.165***(0.001)	-0.164***(0.001)	-0.160***(0.001)	-0.171***(0.001)	-0.171***(0.001)
-PlaceBusiness		-1.305***(0.002)	-1.301***(0.002)	-1.292***(0.002)	-1.294***(0.002)	-1.275***(0.002)
-Websites		0.028***(0.001)	0.038***(0.001)	0.019***(0.001)	0.042***(0.001)	0.041***(0.001)
FRIENDLIKELY			0.072***(0.000)	0.051***(0.001)		0.080***(0.001)
REMFACT				-0.065***(0.001)		-0.021***(0.001)
EMOTION				0.250***(0.001)		0.260***(0.001)
EMOTICON				-0.049***(0.001)		-0.020***(0.001)
HOLIDAYMENTION				-0.187***(0.004)		-0.183***(0.004)
HUMOR				-0.009***(0.000)		0.009***(0.000)
PHILANTHROPIC				-0.072***(0.001)		0.002*(0.001)
SMALLTALK				-0.112***(0.001)		-0.146***(0.001)
BRANDMENTION					0.012***(0.000)	0.021***(0.000)
DEAL					-0.166***(0.001)	-0.207***(0.001)
PRICECOMPARE					-0.036***(0.000)	-0.047***(0.000)
PRICE					-0.504***(0.001)	-0.471***(0.001)
TARGET					0.050***(0.001)	-0.019***(0.001)
PRODAVAIL					-0.027***(0.001)	-0.073***(0.001)
PRODLOCATION					0.134***(0.001)	0.138***(0.001)
PRODMENTION					-0.004***(0.001)	0.012***(0.001)
						Demographic proportion variables excluded
McFadden R-sq.		0.191	0.191	0.195	0.196	0.201
Nagelkerke R-sq.		0.192	0.193	0.197	0.197	0.203
Log-likelihood	-33968732.078	-27584573.527	-27566219.315	-27424516.699	-27419099.696	-27221231.997
Deviance	66990021.316	54221704.214	54184995.789	53901590.557	53890756.552	53495021.154
AIC	67937466.156	55169211.054	55132504.629	54849113.398	54838281.393	54442561.994
N	665916	665916	665916	665916	665916	665916
Significance '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1						

Table 14: **Logistic Regression EdgeRank-Corrected Estimates Model Comparison (*Likes*)**