UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

A Marketing Perspective on Social Media Usefulness

Matthijs Meire

Supervisor: Prof. Dr. Dirk Van den Poel

Academic year: 2017-2018

A dissertation submitted to the Faculty of Economics and Business Administration, Ghent University, in fulfilment of the requirements for the degree of Doctor in Applied Economic Sciences

DOCTORAL JURY

Dean Prof. Dr. Patrick Van Kenhove
(Ghent University)

Prof. Dr. Dirk Van den Poel
(Ghent University)

Prof. Dr. Dries Benoit
(Ghent University)

Prof. Dr. Bart Larivière
(Ghent University)

Prof. Dr. Edward Malthouse
(Northwestern University)

Prof. Dr. Wouter Verbeke
(Vrije Universiteit Brussel)

ACKNOWLEDGEMENTS

Over the past 4 years, I spent my time writing this dissertation. This would not have been possible without the support of many different people who I like to thank here.

First of all, I want to thank my advisor prof. dr. Dirk Van den Poel, who gave me the opportunity to pursue my PhD. He always supported my research ideas, gave valuable feedback and also took into account my family situation. Next, I want to thank Michel Ballings for his support during my PhD. I feel Michel has given invaluable support in the early stages of the dissertation and I was honored to work with him on all of the research projects. Of course I am also grateful to my other co-authors, Kelly Hewett and V. Kumar, for their interesting insights, especially on how to craft and frame research.

Next, I thank all present and past members of the marketing department for the four wonderful years at the department. I will certainly miss the enjoyable environment, the department activities, lunches, soccer games, research discussions and so much more.

I would also like to thank all the jury members, I am very grateful for the time and effort you invested in revising my dissertation.

Finally, I want to thank my family for supporting me and helping me out when necessary. I especially want to thank Ester, Jules and Alixe, for putting things in perspective and giving me so much love during the past 4 years.

Table of contents

List of Tables

List of Figures

NEDERLANDSTALIGE SAMENVATTING

Sociale media omvatten alle internet-gebaseerde applicaties waar gebruikers de inhoud zelf kunnen creëren en delen, en waar ze kunnen interageren met elkaar; de meest gekende voorbeelden zijn Facebook, Twitter, Instagram, Snapchat en blogs. Tegenwoordig worden sociale media ook gebruikt door bedrijven als deel van hun marketing mix, met als voornaamste genoemde voordelen de mogelijkheid om interactief klanten te kunnen engageren en connectie te maken met de klanten. Ondanks deze groeiende interesse in sociale media en de grote investeringen, is de opbrengst van deze investeringen nog steeds onzeker, en academisch onderzoek naar het effect van sociale media marketing hinkt achterop. Hoofdstuk 1 focust op enkele vragen die nog niet ten volle onderzocht werden in de literatuur, legt uit hoe dit doctoraat bijdraagt tot het aanvullen van deze hiaten en toont hoe sociale media verder kan bijdragen tot de het creëren van waarde voor het bedrijf.

In dit doctoraat worden meer specifiek de volgende vragen beantwoord: (1) Hoe kunnen we meer accurate schattingen van sentiment in online commentaren of gesprekken (eWOM) verkrijgen? (2) Welke invloed hebben sociale media en klantensentiment op de waarde van de klant voor een bedrijf?, en (3): Hoe kunnen business-to-business bedrijven sociale media gebruiken binnen de verkoopscyclus?

De drie studies binnen dit doctoraat zijn met elkaar gelinkt doordat ze allen Facebook data gebruiken als belangrijkste informatiebron, en alledrie de analytische toolset voor klantenrelaties en beheer op verschillende niveaus uitbreiden. Voor de eerste studie (Hoofdstuk 2), starten we van gebruikersinformatie (Facebook posts). De tweede studie (Hoofdstuk 3) gebruikt een combinatie van gebruikersinformatie in combinatie met bedrijfsinformatie, in de vorm van Facebook posts die zijn opgesteld door de marketeer (op Facebook pagina's). Ten laatste wordt in studie 3 (Hoofdstuk 4) exclusief de nadruk gelegd op Facebook pagina's en bedrijfsinformatie om een business-to-business predictiesysteem op te zetten.

In Hoofdstuk 2 onderzoeken we hoe automatische sentiment-detectie op sociale media kan worden verbeterd. Sociale media bieden marketeers immers veel mogelijke informatie over klanten, maar het grootste deel van deze informatie is niet gestructureerd (bijvoorbeeld video's, foto's en tekst) waardoor de betekenis op een bepaalde manier moet achterhaald worden. We focussen enkel op tekst in dit hoofdstuk, en meer specifiek op Facebook posts, om het sentiment van deze posts te ontdekken en voorspellen. We starten van een breed basismodel voor sentiment predictie, gebaseerd op de uitgebreid aanwezige literatuur rond dit onderwerp. We

stellen twee alternatieve types extra informatie voor om deze modellen te complementeren. Het eerste type variabelen omvat voorblijvende informatie, waarmee we doelen op informatie die beschikbaar is voordat de echte inhoud gepost is. Voorbeelden van dit type zijn sentiment in eerdere posts en meer algemene gebruikersinformatie zoals bijvoorbeeld demografische informatie. Deze informatie laat ook toe om te kijken naar afwijkingen van normaal post-gedrag om veranderingen in sentiment te detecteren. Het tweede type variabelen omvat achterblijvende informatie, die informatie bevatten die slechts enige tijd na het posten beschikbaar wordt. De meest bekende voorbeelden zijn bijvoorbeeld 'vind-ik-leuks' en 'commentaren' op Facebook, die bijvoorbeeld verzameld kunnen worden na enkele uren/dagen. We delen de informatie op in voorblijvende en achterblijvende informatie, omdat het eerste type kan gebruikt worden in real-time sentiment classificatie, terwijl het laatste type nooit kan gebruikt worden in een real-time setting. Vervolgens bouwen we drie sentiment classificatie modellen, waarbij we 5*2 fold cross-validatie Random Forest modellen gebruiken, om de toegevoegde waarde van voorblijvende en achterblijvende informatie bovenop de basisinformatie te bepalen. De resultaten tonen dat beide soorten informatie waarde toevoegen bovenop het basismodel. Verder wordt duidelijk dat zowel afwijkingen van 'normaal' post gedrag als het aantal 'vind-ik-leuks' en commentaren de performantie van onze modellen substantieel verhogen. We zien ook dat de drie soorten informatie complementair zijn, en allen belangrijk zijn voor de performantie van het meest complete model met alle informatie inbegrepen. Deze resultaten hebben een hoge praktische en academische waarde, aangezien sentiment vaak gebruikt wordt in marketing door de bewezen relatie met verkopen, wat het belangrijk maakt om sentiment correct te meten. Verder kunnen bedrijven ook klanten tevredenheid afleiden uit sociale media sentiment.

Aanraakpunten voor klanten zijn alle momenten waarop klanten in contact kunnen komen met het bedrijf, en omvatten zowel passieve (bv., het bekijken van reclame) als actieve (bv., aankopen) momenten. In hoofdstuk 3 linken we de uitkomsten van zo'n aanraakpunten aan online klantensentiment, gemeten door middel van Facebook commentaar. Verder stellen we voor dat (online) marketeer gegeneerde inhoud die volgt op het specifieke aanraakpunt, een modererend effect heeft op de impact van het resultaat van dit aanraakpunt op het tentoongestelde klantensentiment. Finaal linken we dit klantensentiment aan de klantenwaarde voor het bedrijf, terwijl we controleren voor verschillende variabelen gelinkt aan de interacties tussen klant en bedrijf. Voor dit onderzoek verzamelden we een unieke dataset met een grotere set aan merk gerelateerde sociale media activiteit op klantenniveau dan tevoren,

transactievariabelen op klantenniveau, variabelen die de objectieve performantie van de klanten-aanraakpunten meten en andere marketing variabelen. Door middel van een twee-fasen model waarbij we eerst klantensentiment modelleren in een gegeneraliseerd lineair mixed effect model, gevolgd door een Type II Tobit model voor klantenwaarde, tonen we aan dat marketeer online content klantensentiment kan beïnvloeden na meer negatieve klantenervaringen, en dat klantensentiment direct gerelateerd is aan klantenwaarde zelfs wanneer gecontroleerd wordt voor de andere variabelen. Ten laatste toont dit onderzoek ook aan dat de meest gebruikt item op Facebook, de pagina vind-ik-leuk, geen significant effect heeft op klantenwaarde.

Het overgrote deel van het huidige onderzoek rond sociale media onderzoekt Business-to-Consumer markten, met een focus op de interactiviteit van de conversaties en de potentiële waarde van elektronische commentaren (eWOM). In het laatste hoofdstuk onderzoeken we hoe Business-to-Business bedrijven sociale media kunnen gebruiken in het verkoopsproces. Inderdaad, bedrijven creëren sociale media inhoud, en deze informatie kan vervolgens gebruikt worden door andere bedrijven in hun aankoop(acquisitie)proces. We stellen een klantenacquisitie predictiemodel voor, dat prospecten van een bedrijf kwalificeert als mogelijke klanten. Het model vergelijkt informatie van sociale media (Facebook) profielen van de prospecten met informatie van twee andere databronnen: webpagina informatie en commercieel aangekochte data, en we testen het model met een grootschalig experiment bij Coca Cola Refreshments, Inc. De resultaten tonen aan dat de sociale media informatie het meest informatief is, maar ook dat het complementair is met de informatie van de andere databronnen. Verder toont dit onderzoek aan hoe het modelleren voordeel haalt bij een iteratieve aanpak, en demonstreren we de financiële voordelen van onze voorgestelde aanpak.

Als conclusie kunnen we stellen dat we met dit doctoraat antwoorden hebben kunnen geven op enkele belangrijke vragen met betrekking tot marketing en de interactie ervan met sociale media, waarbij we een significante bijdrage leveren aan zowel theorie als praktijk.

ENGLISH SUMMARY

Social media represent all internet-based applications in which customers can create and share the content, and where they can interact with each other; the most well-known examples are Facebook, Twitter, Instagram, Snapchat and blogs. Nowadays, social media is also used by companies as a part of their marketing mix, with the main advantages named being the possibility to interactively engage with their customers and connect with them. Despite this growing interest in social media and the large investments, the return on these investments is still debated and academic research on the effects of social media marketing is lagging. Chapter 1 focuses on some of the gaps that still exist, explains how this dissertation aims to contribute to literature and shows how social media can contribute to business value.

More specifically, we answer the following three questions in this dissertation: (1) how to provide more accurate estimations of sentiment in online word-of-mouth?, (2) How does social media and customer sentiment impact customer value to the firm? and (3) How can business-to-business (B2B) firms use social media optimally within the sales cycle?

The three studies in this dissertation are related in that they each use Facebook information as the main source of information, and because they extend the analytical toolset available for the management of customer relationships. For the first study (chapter 2), we start from specific user information (Facebook posts). For the second study (chapter 3), we use a combination of individual user information, combined with marketer generated content from Facebook pages (company information). Finally, in study 3 (chapter 4), we exclusively focus on Facebook pages and company information to set up a business-to-business acquisition prediction system.

In chapter 2, we investigate how automatic sentiment detection on social media can be improved. Social media offer a lot of potential for marketers to retrieve information about customers. However, most of this information is unstructured, and it's meaning has to be inferenced in some way. We focus exclusively on textual content, and more precisely Facebook posts, and aim to discover and predict the sentiment of these posts. We start from a broad baseline sentiment classification model, based on the extensively available previous literature, and we suggest two alternative types of extra variables to complement these models. The first type of variables comprises leading information, with which we mean information that is available before the actual content was posted. Examples of this type are sentiment in previous posts and general user information such as demographics. This information also allows to look at deviations from 'normal' posting behavior to detect changes in sentiment. The second type

of variables are lagging variables, which contain information that becomes available only some time after a post has been published. The most noteworthy examples are likes and comments gathered for this post after, for instance, 7 days. We split these information types since leading variables could be used in real-time sentiment classification, while lagging variables will never be real-time. We subsequently build three sentiment classification models, using 5*2 fold cross validation Random Forest models in order to evaluate the added value of the leading and lagging variables. The results show that both leading and lagging variables create significant and relevant value over and above the baseline model. It turns out that deviations from 'normal' posting behavior as well as comments and likes substantially increase our models' performance. We also see that the traditional textual information, leading and lagging information are all complementary and add to model performance in the most complete model. These results have high practical and academic value, since valence is commonly used in marketing as it has a demonstrated relationship with sales, which makes it important to correctly measure valence. Furthermore, consumer sentiment or satisfaction about a brand can be deduced from social media.

Customer touchpoints are all occasions in which customers can relate to a firm, and comprise both passive (e.g., seeing advertisements) and active (e.g., purchasing) moments. In chapter 3, we link the outcome of such customer touchpoints to online customer sentiment measured by Facebook comments. Moreover, we propose that (online) marketer generated content, following the specific touchpoint, can moderate the impact of the result of the touchpoint on the subsequent displayed sentiment. Finally, we link individual customer sentiment to direct engagement (also known as customer lifetime value (CLV)), in combination with several control variables linked to customer-firm interaction data. For this research, we compiled a unique dataset which features an unprecedented set of brand-related customer-level social media activity metrics, transaction variables at the customer level, variables capturing objective performance characteristics of the customer touchpoint and other marketing communication variables. By using a two stage model in which we first model customer sentiment in a generalized linear mixed effects model, followed by a Type II Tobit model for engagement, we show that marketer generated content is able to influence customer sentiment following more negative service encounters, and that customer sentiment is related to direct engagement, even when traditional control variables are included. Finally, this research also shows that the most used Facebook metric, a page like, has no significant effect on direct engagement.

Most of the current research focuses on social media usage for in Business-to-Consumer (B2C) environments, with a focus on the interactivity of conversations and the potential value of electronic word of mouth. In the final chapter, we investigate how Business-to-Business (B2B) organizations can use social media in their sales processes. Indeed, businesses create social media content, and this information can subsequently be used by other companies in their acquisition process. We propose a customer acquisition prediction model, that qualifies a companies' prospects as potential customers. The model compares social media (Facebook) information of the prospect with two other data sources: web page information and commercially purchased information, and we test the model with a large scale experiment at Coca-Cola Refreshments, Inc. The results show that Facebook information is most informative, but that it is complementary to the information from the other data sources. Moreover, this research shows how the modeling efforts can benefit from an iterative approach, and we demonstrate the financial benefits of our newly devised approach.

To summarize, in this dissertation we were able to respond to some relevant and important questions related to marketing and it's interaction with social media, thereby delivering both theoretical and practical contributions.

1

Introduction

Today, many people as well as organizations use social media for communication purposes. One of the earliest definitions of social media is given by Kaplan and Haenlein (2010), who define social media as the internet-based applications that allow the creation and exchange of user generated content. This early definition stresses the fact that originally, most social media tools were developed and used for consumer-to-consumer (C2C) communication only (e.g., blogs, reviews, but also Facebook and Twitter). Social media has, among other evolutions, enabled customers to be no longer passive, but instead be observers, initiators, participants and co-creators (Maslowska et al., 2016). These social media users are even called pseudo-marketers, but with greater influence, lower costs and potentially a more effective reach than actual marketers (Kozinets et al., 2010). Research has shown that some of these C2C-communications have led to positive business outcomes (e.g. Onishi and Manchanda, 2012; Rishika et al., 2013), sparking business interest in these media. Moreover, the digital nature of social media offers firms the possibility to easily track the conversations (Moe and Schweidel, 2017). However, social media contents and its rapid dissemination among the network of consumers can also be negative for brands (Gensler et al., 2013). In order to stay competitive, firms should adapt to the changing environment and embrace social media as a tool to create opportunities and competitive advantage (Hennig-Thurau et al., 2010) and try to manage their brands on social media (Leeflang et al., 2014). It is therefore not surprising that nowadays, social media are frequently seen as part of the marketing mix (Chen and Xie, 2008; John et al., 2017; Mangold and Faulds, 2009), or as a way to get marketing insights (Moe and Schweidel, 2017), and firms aim to integrate social media into customer relationship management (CRM), forming social CRM capabilities (Malthouse et al., 2013; Trainor et al., 2014). Social media

thus have evolved to platforms in which both consumers and firms are present. Firms can not only use social media to reach a wider audience and control brand management, but also to foster engagement and help shaping the entire customer experience. However, entering social media as a firm still entails several pitfalls, such as measuring the ROI of social media investments, lack of control and insight in message diffusion and difficulties to integrate customer touch points (Malthouse et al., 2013).

Since social media is still a relatively new and continuously evolving marketing tool, research about its (potential) impact on consumers and companies is still scarce, and not in proportion to the business social media focus and budgets that are spend on social media. Indeed, recent estimates of global social media marketing in 2017 are as high as 13.5 billion US dollars, and even more when digital advertising is taken into account (Statista, 2018). Moreover, a recent analysis showed that the wide majority (98%) of the Fortune 500 companies are active on social media (Ganim Barnes and Pavao, 2017). Despite these investments, in-depth knowledge about many of the aspects of social media, and especially their business value, is still lacking. Several critical research questions with practical relevance still remain unanswered, that all relate to the business value of social media, such as (1) how to provide more accurate classifications of sentiment in online word-of-mouth?, (2) How do social media and customer sentiment impact customer value to the firm? and (3) How can business-to-business (B2B) firms use social media optimally within the sales cycle? In the following paragraphs, we briefly outline why these questions are important and how this dissertation aims to answer them. At the same time, we outline the most relevant social media literature related to each of these questions.

1. How to provide more accurate classifications of sentiment in online word of mouth?

Most research relating to social media considers electronic word of mouth (eWOM), also called user-generated content (UGC) or consumer-generated content. This seems logical, given that social media was established as a C2C communication tool. Several studies link eWOM to different subsequent behavioral outcomes such as sales. Babić Rosario et al. (2015) provide a recent meta-analysis of eWOM-related papers in several domains, in which they show the positive effect on sales, but also show that the effectiveness differs based on online platforms, products, and eWOM metrics. Their study shows that with regard to the metrics, volume has a stronger impact on sales than valence (e.g., percentage of negative ratings) and a similar impact

as a composite volume-valence metric (e.g., the volume of positive or negative ratings). Moreover, they find that negative eWOM not necessarily leads to lower sales, but a high variation in positive and negative eWOM does (Babić Rosario et al., 2016). In summary, these findings argue that both volume and valence based metrics have an impact on sales, which makes it important to accurately estimate these metrics. Since eWOM volume is a relatively straightforward measure, we focus on valence and propose a new approach to make better predictions about eWOM valence. Next to the link with sales, online valence is used in a wide range of other applications. Examples include online valence to inform a company about the overall sentiment with regard to a brand or brand perceptions (Smith et al., 2012; Schweidel and Moe, 2014; Tirunillai and Tellis, 2014), predicting election outcomes (Tumasjan et al., 2010) or increasing online learning performance (Ortigosa et al., 2014). Enhancing valence prediction thus offers possibilities to increase system performance for many applications.

Since the rise of social media and eWOM, there has been a growing interest in valence prediction. While part of the (marketing) research uses straightforward tools such as star ratings or sentiment dictionaries to determine valence, there has been a growing body of literature in which the valence prediction itself became the main subject of interest. Going beyond the basic star ratings allows to include more subtle textual information that is given in an eWOM instance (Archak et al., 2011). This stream of literature is now called 'sentiment analysis', which is defined as the computational process of extracting sentiment from text and has clear influences from text analysis and machine learning techniques (e.g, Liu, 2012; Pang et al., 2002 for early examples). The literature has discussed a myriad of possibilities to improve sentiment classification based on better machine learning models or better features. However, the features proposed are still mainly based on the textual characteristics of the eWOM instance (a post, a review, …), without taking into account the characteristics of the reviewer/poster and other available information. This is surprising, given that in the related field of review helpfulness, reviewer characteristics have long been included (e.g., Forman et al., 2008; Ghose and Ipeirotis, 2011). This shows that past reviewer behavior and information are very informative for review helpfulness prediction (Ghose and Ipeirotis, 2011), and thus may also help to increase valence prediction. Other studies have shown that crowd-based information, which becomes available after posting, also helps to increase prediction performance (Hoornaert et al., 2017). Therefore, in Chapter 2, we set out to improve existing sentiment analysis models by using previously untapped information (Meire et al., 2016). More specifically, we analyze the added value of leading and lagging auxiliary information to assess or classify sentiment of Facebook posts.

Leading information in this context is defined as content that is already available when a post or comment is placed on Facebook. This includes general user information as well as previous post information. This information is similar to the information taken into account by Ghose and Ipeirotis (2011), who include reviewer history in an online rating environment. Lagging information consists of all information that only becomes available some time after a post or comment is put on Facebook. The main examples of lagging variables include likes and comments on the post. We subsequently compare the performance of a baseline prediction model, which includes the typical textual sentiment analysis features, and models that include leading and lagging information. This paper is the first to take into account leading and lagging information, and proves that sentiment classifications can be substantially improved by taking this extra information into account. This approach thus proves relevant for both research and practice, and for both real-time and delayed sentiment classification.

2. How do social media and customer sentiment impact customer value to the firm?

Several authors have already specified the need, and also the difficulty, to measure the return on investment of social media endeavors (e.g., Malthouse et al., 2013). It is thus not surprising that an increasingly large body of literature investigates the value of social media. In this research, typically the value of eWOM is evaluated in a longitudinal fashion with VAR-related models (see e.g. Babić Rosario et al. (2015) for an overview). More recent papers also take into account the interaction effects between eWOM, marketer generated content (MGC) and more traditional marketing activities such as advertising (e.g. Colicev et al., 2018; de Vries et al., 2017; Hewett et al., 2016; Manchanda et al., 2015), and show that these different aspects operate within an echoverse, reverberating the effects of one another (Hewett et al., 2016). It makes clear that social media are not a standalone marketing tool, but rather part of the broader communication or marketing mix. However, while making great progress in our understanding of social media working mechanisms, these studies share some limitations, and several research questions remain unanswered.

First, these papers focus on firm or product level outcomes such as sales, stock prices or brand awareness. Results relating social media activity to value on an individual level is more scarce, although research is catching up with recent studies looking at the value of sending and receiving MGC and UGC on Facebook and the value of liking a Facebook fan page (Goh et al., 2013; Mochon et al., 2017; Xie and Lee, 2015; Malthouse et al., 2016). Second, MGC and UGC

are mostly limited to single measures, on the firm or individual level, such as the volume or valence of Tweets or Facebook posts on a global level, or the act of liking a Facebook page on an individual level (John et al., 2017; Kumar et al., 2016; Mochon et al., 2017), without integrating multiple aspects of MGC and UGC. Moreover, studies investigating network effects and the economic value of these networks on social media is very scarce (e.g., Zhang and Pennacchiotti, 2013), while social networks are especially designed around these networks of people who can interact with and influence a focal customer. Third, there is little research linking social media usage to specific customer touchpoints, whether they are offline or online. This is surprising, given the spike of interest in concepts such as the customer journey, the customer touch points on this journey (Homburg et al., 2015; Schmitt, 2003) and the customer experience with these touch points on the one hand (Lemon and Verhoef, 2016), and the marketer's ability to monitor objective performance criteria related to these experiences (e.g., store traffic patterns, waiting times, etc.) on the other hand. Moreover, social media also offer the possibility to continuously and in real-time track user generated comments related to the firm, which are already shown to capture brand perceptions (Schweidel and Moe, 2014) or predict purchase behavior (Baker et al., 2016). Current research has aggregated MGC or UGC mostly in time intervals (weeks or months), or looked at specific UGC moments of interest (e.g., liking a page), instead of integrating social media MGC and UGC around particular touch points or customer experience encounters. Hence, these studies are not able to link UGC or MGC to specific experiences or to evaluate how the customer's sentiment related to these experiences can be captured or influenced by MGC actions. However, recent work by Harmeling et al. (2017) argues that marketer's actions could enhance the effect of customers' experience on the customers' value to the firm. We thus view marketers' abilities to influence sentiment and drive customers' value as a result of customer experiences remains an un-tapped use of social media and an under-researched area in the marketing domain.

Therefore, in chapter 3 we set out to fill the gaps identified above. We link customer experience, related to specific customer experience encounters and measured by objective performance criteria, to (online) individual customer sentiment regarding these encounters in a soccer team context. Moreover, we propose that (online) marketer generated content, following the specific encounters, can moderate the impact of the result of the encounter on the subsequent displayed sentiment. Finally, we link individual customer sentiment to direct engagement (also known as customer lifetime value (CLV)), in combination with several control variables linked to customer-firm interaction data. We do this by collecting a unique dataset; it features an

unprecedented set of brand-related customer-level social media activity metrics including liking a brand's social media page, MGC, likes of and comments on the brand's posts, and RSVPs for events sponsored by the brand. In addition, our data include transaction variables at the customer level, variables capturing objective performance characteristics of the experience encounter and other marketing communication variables. The results show that marketer generated content is able to influence customer sentiment following more negative experience encounters, and that customer sentiment is related to direct engagement, even when traditional control variables are included. Finally, we note that the most used Facebook metric, a page like, has no significant effect on direct engagement.

3. How can business-to-business (B2B) firms optimally use social media within the sales cycle?

Social media is typically seen as a useful tool in Business-to-Consumer (B2C) marketing domains. Although there is practical evidence about the importance of social media for Business-to-Business (B2B) marketing (e.g., Gillin and Schwartzman, 2010; Shih, 2010, Wang et al., 2017), most research on social media still focuses on B2C domain and the adoption of social media by B2B companies has been slow (Michaelidou et al., 2011). The internal use of social media is relatively higher compared to the external use (e.g., with partners or organizations) (Jussila et al., 2014), and it has been recognized that social media can foster communication, interaction, learning and communication among employees (e.g. García-Peñalvo et al., 2012).

The use of external social media is mostly related to the sales and marketing. Jussila et al. (2014), for example, discuss the possibilities of employer branding, general communication and sales support. Trainor et al. (2014) focus on the valuable information regarding customer requirements, complaints and experiences that can be gained from social media, and Ustüner and Godes (2006) mention an increase in effectiveness driven by a better understanding of the underlying social networks between customers and prospects. Some scholars focus on social media as 'content enabler' during the sales process (Agnihotri et al., 2012; Järvinen and Taiminen, 2016; Rodriguez et al., 2012). However, most research focuses on the different stages of the sales process and the role social media can play in these stages. Michaelidou et al. (2011) mention that social media are valuable for attracting new customers, cultivating relationships and supporting brands. Giamanco and Gregoire (2012) suggest three stages in which social media can be used, prospecting (i.e., finding new leads), qualifying leads and managing

relationships. Similarly, Rodriguez et al. (2012) identified 3 steps: creating opportunity, understanding the customers and relationship management. It is clear that these steps largely coincide with the ones defined by Giamanco and Gregoire (2012). Finally, several researches have focused on the entire sales process in more detail (e.g. Agnihotri et al., 2016; Andzulis et al., 2012; Marshall et al., 2012) to also include building trust and customer service.

Most of the research above mentions tools such as LinkedIn and Facebook as the main social media applications that can be used. They posit ideas and frameworks and elaborate on how salespeople can identify new prospects, on how they can use social media to identify the good prospects and how social media can be used to start or maintain the relationship with the customer. Social media are recognized as a tool to make the sales process less costly and more effective and is seen as an extension of traditional CRM, leading to Social CRM activities (Trainor et al., 2014). Although sharing the common idea that social media are important in a B2B selling context, only one paper provides evidence that social media has a positive relationship with selling organizations' ability to both create opportunities and manage relationships (Rodriguez et al., 2012). This same study also states that the sales team performance (e.g., number of identified prospects, number of acquired new customers) is positively influenced by social media usage (Rodriguez et al., 2012).

The studies concerning social media use in the sales cycle thus share one common shortcoming, namely that they are mostly explorative in nature and do not relate the potential of social media to proven business advantages. The only study that does so (Rodriguez et al., 2012), is also limited in that they rely on questionnaire data related to the value of social media. We feel that this current qualitative focus on social media in the literature ignores important opportunities, related to the big data nature of social media. Social media offer opportunities for automatic extraction and processing of data and the use of advanced analytical techniques to gain insights from these data (Lilien, 2016). Therefore, in Chapter 4, we aim to overcome these limitations and take a different view on social media in the sales process by looking at social media as big data (Meire et al., 2017). Instead of seeing social media mainly as a communication tool, we use social media data to create a customer acquisition support system to qualify prospects as potential customers. More specifically, we evaluate the predictive value of data extracted from the prospects' social media page (Facebook pages), and compare this with data extracted from their website and commercial data bought from a specialized vendor in a real-life experiment with Coca-Cola Refreshments USA. In this way, me make three major contributions to literature: First, we posit, evaluate and assess a customer acquisition system on

a large scale and show the financial benefits of this approach. Second, we add to the existing B2B social media literature by taking a quantitative view. Third, we add to existing literature on B2B acquisition by incorporating a new, freely available data source and show that this is the most important information source for prospect conversion prediction. With this research, we not only provide insight into the value of actual social media data, we also respond to recent calls for more B2B customer analytics, driven by new data sources such as social media activity.

4. Overview

In sum, with this dissertation, we aim to contribute to literature by investigating the added value of social media for creating business value. Hence, we can position this dissertation within a social media and within a business value (CRM) framework.

4.1. Social media framing

We focus on Facebook as a social media platform, as shown in Figure 1. Research using Facebook data is scarce as most research related to social media focuses on blog posts, reviews or Twitter in a B2C environment. This might seem surprising, as Facebook is the biggest and most widely used social media platform nowadays (John et al., 2017), allowing both firms and consumers to join the network (unlike e.g. blog posts) and providing a wide range of interaction possibilities (events, likes, comments, shares, posts, reviews, etc.). Facebook is also the easiest tool to connect social media users to actual customers once Facebook data is available. It is, however, more difficult to collect (individual user) Facebook information using a public API as Facebook is more closed compared to for instance Twitter. Twitter allows to extract profile information of every user, and the majority of the users keeps their tweets open to the public. Facebook, in contrast, has strict privacy regulations that keep on getting tighter (e.g., Constine, 2015).

Figure 1.1 shows the main elements of Facebook, a user profile and a fan page (e.g., a company's brand fan page), and how these elements are incorporated into this dissertation. We consider both Facebook user pages information (sentiment analysis in Chapter 2), elements regarding the use of companies' Facebook fan pages (Chapter 4) and a combination of user specific information and Facebook fan page information in Chapter 3 to create a comprehensive view of the value of Facebook as a business tool. For Chapter 2 and 3, we rely on social media data that was extracted using a Facebook app in 2014, and augmented with the fan page data for Chapter 3. For Chapter 4, we rely exclusively on publicly available Facebook fan pages.

4.2. CRM framing

The chapters are not only related in that they focus on social media (Facebook), they can also be seen within an analytical approach to create business value. We illustrate this in Figure 1.2. We start from the central concept of Customer Relationship Management (CRM). This concept entails all relationship efforts between a company and (would be) customers, mostly initiated

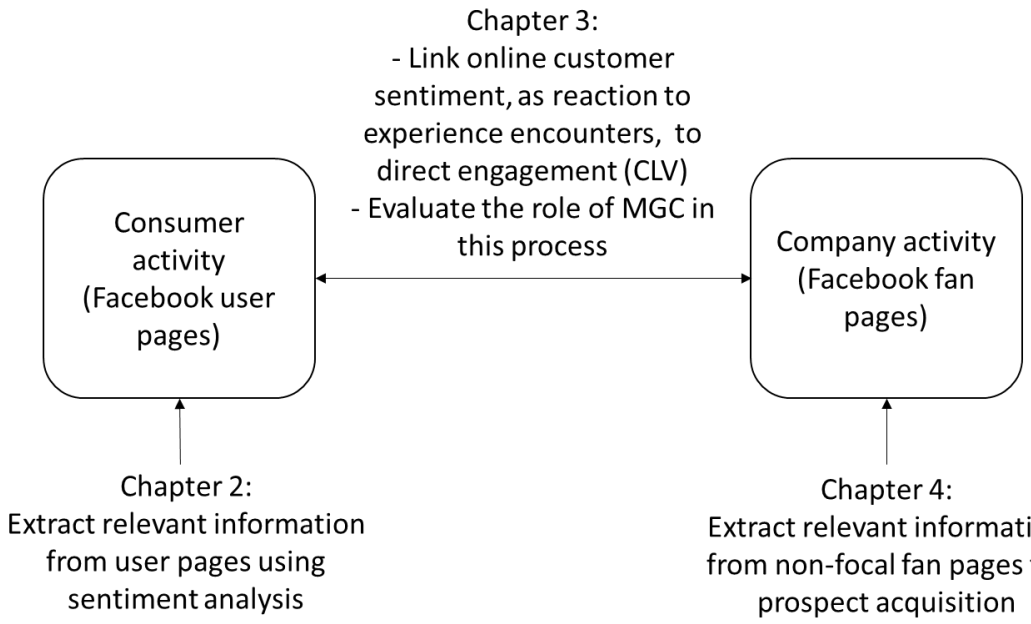


Figure 1.1: Graphical overview of this dissertation from a social media perspective

by the company, with the ultimate goal of maximizing CLV. Typical CRM activities are depicted in Figure 1.2, and comprise customer acquisition efforts, customer up-sell and cross-sell and customer retention management. More recent insights mention that CRM can be broadened, to capture the entire customer journey with the firm (Lemon and Verhoef, 2016), as a consequence of the emergence of 'empowered' customers and UGC (Edelman & Singer, 2015). The chapters discuss several innovative applications that can enrich the analytical toolset for CRM or the customer journey, which will eventually lead to better business objectives. In Chapter 2, we focus on new methods regarding customer sentiment, which can be applied to the entire customer journey. In Chapter 3, we discuss the impact of social media on CLV, using advanced econometric models. Finally, Chapter 4 narrows down the scope to customer acquisition, focusing on the added value of social media in a B2B context. Thus, by investigating the relatively new data source of social media and its inclusion in a CRM context, we broaden the arsenal of data analytical tools for CRM.

All individual chapters are based on academic articles published in or ready to submit to academic journals, and can also be read separately. Table 1.1 further summarizes the chapters of this dissertation, highlighting their contributions, key results, the type of research, applied methods and data sources.

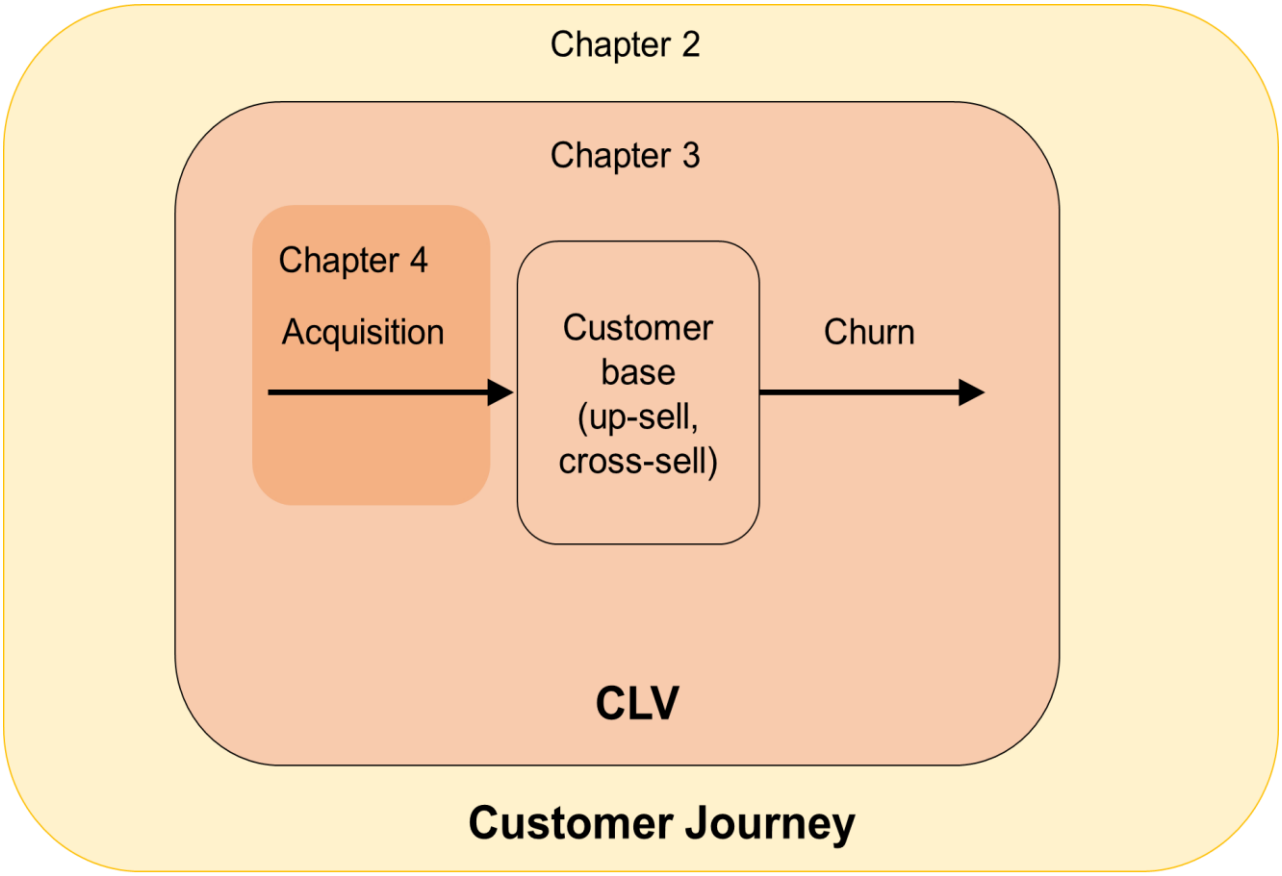


Figure 1.2: Overview of the dissertation from a CRM/macustomer journey perspective

Chapter	Contribution	Key Results	Type of Research	Applied methods	Data sources
2. The added value of auxiliary data in sentiment analysis of Facebook posts	Show the added value of leading and lagging information for sentiment analysis Provide framework for sentiment analysis	Leading and lagging information add predictive power Leading, lagging and focal post information are complementary Likes and comments on posts are very informative for the sentiment of that post	Predictive	Random Forests, Support Vector Machines	- Social media posts (Facebook)
3. Linking Customer Experience to Customer Engagement: the Role of MGC and Customer Sentiment	Clarify the role of MGC in shaping customer sentiment as a result of actual customer experience encounters Link online customer sentiment to customer value (direct engagement)	We find that the volume of MGC has a moderating effect on the objective influence of a service encounter outcome on customer sentiment Customer sentiment is positively related to direct engagement (CLV), with a stronger influence on purchase likelihood than on contribution margin Page likes are not significantly related to direct engagement	Descriptive	Random effect regression Type II Tobit model with random effects	- Social media data (Facebook activity metrics on the fan page) - Transaction data on customer and network level - Marketing communication variables - Objective experience characteristics
4. The added value of social media data in B2B customer acquisition systems: a real-life experiment	Posit and assess an acquisition support system on a large scale and show the financial benefits Take a quantitative view on social media Incorporate social media in customer acquisition models for better prediction models	Social media data have higher predictive ability compared to commercial and website data, thanks to richer data Social media data is complementary with the other data sources Phased models are beneficial for customer acquisition models Large potential financial benefits	Predictive	Random Forests with profitability analysis	Prospect and customer data: - Social media data (Facebook pages) - Website data - Commercial vendor data Transaction data (Coca-Cola Refreshments USA)

Table 1.1: Overview of studies

5. References

Agnihotri, R., Dingus, R., Hu, M.Y., Krush, M.T., 2016. Social media: Influencing customer satisfaction in B2B sales. Industrial Marketing Management 53, 172–180.

Agnihotri, R., Kothandaraman, P., Kashyap, R., Singh, R., 2012. Bringing "Social" Into Sales: The Impact of Salespeople'S Social Media Use on Service Behaviors and Value Creation. Journal of Personal Selling & Sales Management 32, 333–348.

Andzulis, J. "Mick," Panagopoulos, N.G., Rapp, A., 2012. A Review of Social Media and Implications for the Sales Process. Journal of Personal Selling & Sales Management 32, 305–316.

Archak, N., Ghose, A., Ipeirotis, P.G., 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Management Science 57, 1485–1509.

Babić Rosario, A., Sotgiu, F., De Valck, K., Bijmolt, T.H.A., 2016. The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. Journal of Marketing Research 53, 297–318.

Baker, A.M., Donthu, N., Kumar, V., 2016. Investigating How Word-of-Mouth Conversations About Brands Influence Purchase and Retransmission Intentions. Journal of Marketing Research 53, 225–239.

Chen, Y., Xie, J., 2008. Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix. Management Science 54, 477–491.

Colicev, A., Malshe, A., Pauwels, K., O'Connor, P., 2018. Improving Consumer Mindset Metrics and Shareholder Value Through Social Media: The Different Roles of Owned and Earned Media. Journal of Marketing 82, 37–56.

Constine, J., 2015. Facebook Is Shutting Down Its API For Giving Your Friends' Data To Apps. TechCrunch. URL http://social.techcrunch.com/2015/04/28/facebook-api-shut-down/ (accessed 10.04.2017)

de Vries, L., Gensler, S., Leeflang, P.S.H., 2017. Effects of Traditional Advertising and Social Messages on Brand-Building Metrics and Customer Acquisition. Journal of Marketing 81, 1–15.

Edelman, D.C., Singer, M., 2015. Competing on Customer Journeys. Harvard Business Review 93, 88–100.

Forman, C., Ghose, A., Wiesenfeld, B., 2008. Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. Information Systems Research 19, 291–313.

Ganim Barnes, N., Pavao, S., 2017. 2017 Fortune 500 - UMass Dartmouth [WWW Document]. URL http://www.umassd.edu/cmr/socialmediaresearch/2017fortune500/#d.en.963986 (accessed 1.5.18).

García-Peñalvo, F.J., Colomo-Palacios, R., Lytras, M.D., 2012. Informal learning in work environments: training with the Social Web in the workplace. Behaviour & Information Technology 31, 753–755.

Gensler, S., Völckner, F., Liu-Thompkins, Y., Wiertz, C., 2013. Managing Brands in the Social Media Environment. Journal of Interactive Marketing, Social Media and Marketing 27, 242–256.

Ghose, A., Ipeirotis, P.G., 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Transactions on Knowledge and Data Engineering 23, 1498–1512.

Giamanco, B., Gregoire, K., 2012. Tweet Me, Friend Me, Make Me Buy. Harvard Business Review 90, 88–93.

Gillin, P., Schwartzman, E., 2010. Social Marketing to the Business Customer: Listen to Your B2B Market, Generate Major Account Leads, and Build Client Relationships. John Wiley & Sons.

Goh, K.-Y., Heng, C.-S., Lin, Z., 2013. Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research 24, 88–107.

Harmeling, C.M., Moffett, J.W., Arnold, M.J., Carlson, B.D., 2017. Toward a theory of customer engagement marketing. J. of the Acad. Mark. Sci. 45, 312–335.

Hennig-Thurau, T., Malthouse, E.C., Friege, C., Gensler, S., Lobschat, L., Rangaswamy, A., Skiera, B., 2010. The Impact of New Media on Customer Relationships. Journal of Service Research 13, 311–330.

Hewett, K., Rand, W., Rust, R.T., van Heerde, H.J., 2016. Brand Buzz in the Echoverse. Journal of Marketing 80, 1–24.

Homburg, C., Ehm, L., Artz, M., 2015. Measuring and Managing Consumer Sentiment in an Online Community Environment. Journal of Marketing Research 52, 629–641.

Hoornaert, S., Ballings, M., Malthouse, E. C., & Van den Poel, D. (2017). Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time. Journal of Product Innovation Management, 34(5), 580-597.

Järvinen, J., Taiminen, H., 2016. Harnessing marketing automation for B2B content marketing. Industrial Marketing Management 54, 164–175.

John, L.K., Emrich, O., Gupta, S., Norton, M.I., 2017. Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes. Journal of Marketing Research 54, 144–155.

Jussila, J.J., Kärkkäinen, H., Aramo-Immonen, H., 2014. Social media utilization in business-to-business relationships of technology industry firms. Computers in Human Behavior 30, 606–613.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 53, 59–68.

Kozinets, R.V., de Valck, K., Wojnicki, A.C., Wilner, S.J.., 2010. Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities. Journal of Marketing 74, 71–89.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., Kannan, P. k., 2016. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. Journal of Marketing 80, 7–25.

Leeflang, P.S.H., Verhoef, P.C., Dahlström, P., Freundt, T., 2014. Challenges and solutions for marketing in a digital era. European Management Journal 32, 1–12.

Lemon, K.N., Verhoef, P.C., 2016. Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing 80, 69–96.

Lilien, G.L., 2016. The B2B Knowledge Gap. International Journal of Research in Marketing 33, 543–556.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 5, 1–167.

Malthouse, E.C., Haenlein, M., Skiera, B., Wege, E., Zhang, M., 2013. Managing Customer Relationships in the Social Media Era: Introducing the Social CRM House. Journal of Interactive Marketing, Social Media and Marketing 27, 270–280.

Malthouse, E. C., Calder, B. J., Kim, S. J., & Vandenbosch, M. (2016). Evidence that user-generated content that produces engagement increases purchase behaviours. Journal of Marketing Management, 32(5-6), 427-444.

Manchanda, P., Packard, G., Pattabhiramaiah, A., 2015. Social Dollars: The Economic Impact of Customer Participation in a Firm-Sponsored Online Customer Community. Marketing Science 34, 367–387.

Mangold, W.G., Faulds, D.J., 2009. Social media: The new hybrid element of the promotion mix. Business Horizons 52, 357–365.

Marshall, G.W., Moncrief, W.C., Rudd, J.M., Lee, N., 2012. Revolution in Sales: The Impact of Social Media and Related Technology on the Selling Environment. Journal of Personal Selling & Sales Management 32, 349–363.

Maslowska, E., Malthouse, E.C., Collinger, T., 2016. The customer engagement ecosystem. Journal of Marketing Management 32, 469–501.

Meire, M., Ballings, M., Van den Poel, D., 2017. The added value of social media data in B2B customer acquisition systems: A real-life experiment. Decision Support Systems.

Meire, M., Ballings, M., Van den Poel, D., 2016. The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems 89, 98–112.

Michaelidou, N., Siamagka, N.-T., Christodoulides, G., 2011. Usage, barriers and measurement of social media marketing: an exploratory investigation of small and medium B2B brands. Industrial Marketing Management 40, 1153–1159.

Mochon, D., Johnson, K., Schwartz, J., Ariely, D., 2017. What Are Likes Worth? A Facebook Page Field Experiment. Journal of Marketing Research 54, 306–317.

Moe, W.W., Schweidel, D.A., 2017: Opportunities for Innovation in Social Media Analytics. Journal of Product Innovation Management 34 (5) 697-702

Onishi, H., Manchanda, P., 2012. Marketing activity, blogging and sales. International Journal of Research in Marketing 29, 221–234.

Ortigosa, A., Martín, J.M., Carro, R.M., 2014. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior 31, 527–541.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 79–86.

Rishika, R., Kumar, A., Janakiraman, R., Bezawada, R., 2013. The Effect of Customers' Social Media Participation on Customer Visit Frequency and Profitability: An Empirical Investigation. Information Systems Research 24, 108–127.

Rodriguez, M., Peterson, R.M., Krishnan, V., 2012. Social Media's Influence on Business-to-Business Sales Performance. Journal of Personal Selling & Sales Management 32, 365–378.

Schmitt, B.H., 2003. Customer Experience Management: A Revolutionary Approach to Connecting with Your Customers, 1 edition. ed. Wiley, New York.

Schweidel, D.A., Moe, W.W., 2014. Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. Journal of Marketing Research 51, 387–402.

Shih, C., 2010. The Facebook Era: Tapping Online Social Networks to Market, Sell, and Innovate, 2 edition. ed. Addison-Wesley Professional, Upper Saddle River, NJ.

Smith, A.N., Fischer, E., Yongjian, C., 2012. How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? Journal of Interactive Marketing 26, 102–113.

Statista, 2018. Social media marketing spending in the U.S. 2017 [WWW Document]. URL https://www.statista.com/statistics/276890/social-media-marketing-expenditure-in-the-united-states/ (accessed 1.5.18).

Tirunillai, S., Tellis, G.J., 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research 51, 463–479.

Trainor, K.J., Andzulis, J. (Mick), Rapp, A., Agnihotri, R., 2014. Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. Journal of Business Research 67, 1201–1208.

Tumasjan, A., Springer, T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 178–185.

Ustüner, T., Godes, D., 2006. Better sales networks. Harv Bus Rev 84, 102–112, 188.

Wang, W. L., Malthouse, E. C., Calder, B., & Uzunoglu, E. (2017). B2B content marketing for professional services: In-person versus digital contacts. Industrial Marketing Management.

Xie, K., Lee, Y.-J., 2015. Social Media and Brand Purchase: Quantifying the Effects of Exposures to Earned and Owned Social Media Activities in a Two-Stage Decision Making Model. Journal of Management Information Systems 32, 204–238.

Zhang, Y., Pennacchiotti, M., 2013. Predicting purchase behaviors from social media, in: Proceedings of the 22nd International Conference on World Wide Web. WWW '13. Edited by: Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo A. Baeza-Yates, and Sue B. Moon, pp. 1521–1532.

2

The Added Value of Auxiliary Data in Sentiment Analysis of Facebook Posts

2. The Added Value of Auxiliary Data in Sentiment Analysis of Facebook Posts

Abstract

The purpose of this study is to (1) assess the added value of information available before (i.e., leading) and after (i.e., lagging) the focal post's creation time in sentiment analysis of Facebook posts, (2) determine which predictors are most important, and (3) investigate the relationship between top predictors and sentiment. We build a sentiment prediction model, including leading information, lagging information, and traditional post variables. We benchmark Random Forest and Support Vector Machines using five times twofold cross validation. The results indicate that both leading and lagging information increase the model's predictive performance. The most important predictors include the number of uppercase letters, the number of likes and the number of negative comments. A higher number of uppercase letters and likes increases the likelihood of a positive post, while a higher number of comments increases the likelihood of a negative post. The main contribution of this study is that it is the first to assess the added value of leading and lagging information in the context of sentiment analysis.

This chapter is based on the published article Meire, M., Ballings, M., Van den Poel, D., 2016. The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems 89, 98–112.

1. Introduction

In the beginning of the century, Web 2.0 emerged as an ideological and technical foundation giving rise to the massive production of user generated-content (UGC). Blogging platforms and online retailers are the first examples of this foundation (Kaplan and Haenlein, 2010). Today, UGC is still growing rapidly, sparking interest and activity in opinion mining and sentiment analysis (Martinez-Camara et al.,2014; Pang and Lee, 2008). Sentiment analysis is defined as the computational process of extracting sentiment from text (Liu, 2012; Pang and Lee, 2008). Applications range from the prediction of election outcomes (Bollen et al., 2009; Tumasjan et al. 2010), to relating public mood to socio-economic variables (Bollen et al., 2009), to improved e-learning strategies (Ortigosa et al., 2014b).

Early examples of sentiment analysis were mainly based on review data. This type of data rarely contained much more information than the content and the time of posting of the review itself. Models using these data are based on present information, where 'present' refers to the time of posting. This changed with the advent of social networks such as Facebook and Twitter in that much more data became available. On these platforms, not only the focal post's content is available, but, taking into account the time of posting, there is also leading and lagging information. Leading information is available even before content is posted (e.g., user profiles, previous posts) and thus contains information about the past. On the other hand, lagging information is generated a posteriori, after the content was posted (e.g., interactions such as likes or retweets) and thus contains information about the future (seen from the time of posting). Leading information can therefore be included in any sentiment model, while lagging information can be included in tools that do not require real-time sentiment analysis. To the best of our knowledge, there is no study that includes leading and lagging information into sentiment analysis models. However, we believe that we can improve sentiment prediction by including leading and lagging information for several reasons. First, social media suffer from a lot of slang (Go et al., 2009; Ortigosa et al., 2014b) making it harder for traditional methods to achieve satisfactory model performance on text variables alone. Second, leading variables would take into account users' average sentiment, word use, well-being, and mood and demographics, effectively acting as a user-specific informative prior of future sentiment and accounting for heterogeneity among users. Leading variables have been shown to lead to better predictions (Basiri et al., 2014). Third, extant literature has found significant relationships between post sentiment and lagging information such as likes, comments and retweets (Stieglitz and Dang-Xuan, 2012; Stieglitz and Dang-Xuan, 2013).

To fill this gap in literature, we assess the additional value for sentiment analysis of leading and lagging information over and above information extracted from the focal post. We do this by constructing three models. The first model is the base model that focuses on the present and contains only the focal post (including text and timing of posting). The second model contains both the focal post's content and leading information, and thus contains both present and past information. Finally, the third model augments the second model with lagging information. This means that the third model takes into account the past, present and future information of a post.

The remainder of this article is structured as follows. First, we provide a literature review focusing on sentiment analysis of social media data and the reasons why leading and lagging information might be valuable in a sentiment prediction model. Second, we detail our methodology including the data, the model description, the predictors, the predictive algorithms and the model evaluation measure. The third section discusses the results. The penultimate section consists of the conclusion and practical implications of this research. In the final section we address the limitations and avenues for future research.

2. Literature review

There are two main approaches to sentiment analysis (Ortigosa et al., 2014b; Taboada et al., 2001). The first approach consists of lexicon-based models, which use predefined lexicons of positive, neutral and negative words to assign positivity values to a sentence or text (e.g., Hatzivassiloglou and Wiebe, 2000; Turney, 2002). Machine learning-based methods constitute the second approach. These methods use several text features (e.g., syntactic features and lexical features; we refer to McInnes (2009) for a complete overview of these features) as input for a training model and predict the sentiment of text using these features (Taboada et al., 2011). Machine learning methods have been shown to be more accurate than lexicon-based methods in general, but also more time consuming (Chaovalit and Zhou, 2005; Pang et al., 2002). Lexicon based methods, however, tend to perform better in less-bounded domains (Ortigosa et al., 2014b). Recently, the two approaches have been combined by several authors (Li and Wu, 2010; Melville et al. 2009; Ortigosa et al., 2014b; Tan et al. 2008; Zhang et al., 2011), mostly by using the scores from a lexicon-based exercise as input features for the machine learning algorithm. In this study we will adopt such a hybrid approach. The reason is that the approach allows for additional features to be added to the model.

Literature on sentiment analysis can be summarized according to (1) the use of a focal post's features (McInnes, 2009), (2) the use of auxiliary features (Basiri et al., 2014), and (3) the focal post's source (Abbasi et al., 2008). The focal post's features constitute: (1) lexicon features, which denote either a pure lexicon based approach or a combination of lexicon and machine learning, (2) lexical features (bag-of-words, n-grams, co-occurrence and collocations), (3) syntactic features (morphology, part-of-speech) and (4) time features. The auxiliary features are divided into leading and lagging features. The former denotes all the information, with regard to a specific user, that is available until the moment of posting. The latter includes information that is available one week after posting (i.e., information on the likes and the comments a post has received). Stated differently, the focal post's features reflect all information of the present, where 'the present' refers to the time of posting, which will be different for every post. Every action that occurred before the present, is referred to as 'the past', while 'the future' indicates all actions that occurred after posting. The leading variables thus originate in the past, while the lagging variables originate in the future.

Table 2.1 provides a representative overview of literature with a focus on social media applications, as social media contain leading and lagging information. It is apparent that sentiment analysis has been widely applied to a diverse set of social media. Table 1 shows that both the lexicon-based (denoted an x in the column labeled 'Lexicon') and the machine learning approaches have been used, and that plenty of text features have been explored. However, it also shows that there is a large potential source of information for sentiment analysis that remains largely untapped. Indeed, social media do not only offer an efficient way to gather the focal post's textual data used in traditional sentiment analysis, they also allow to gather a lot of auxiliary data (e.g., user profile information, likes on statuses) that have not yet been used in sentiment analysis. Basiri et al. (2014) recently made an effort to incorporate such data into a sentiment analysis model. They found that deviations of a reviewer's post compared to the previous posts of this same reviewer lead to better review score prediction. The model of Basiri et al. (2010) is, however, limited to the incorporation of one auxiliary variable and therefore does not reflect the full potential. Furthermore, they do not incorporate the leading information into a sentiment analysis model, but only use it for the prediction of review scores. In this study we will exploit the focal post's information as well as auxiliary leading and lagging data that are present on Facebook. This allows us to assess the improvement in the prediction of emotional valence of Facebook statuses that stems from incorporating auxiliary data. The following section clarifies why leading and lagging information may be important (i.e., improve

the predictive performance of our models). This information is also summarized in the conceptual framework depicted in Fig. 2.1.

Table 2.1: Literature overview

	Features of the focal post				Auxiliary features		Text source
	Lexicon	Lexical	Syntactic	Time	Leading	Lagging	
Pang et al. (2002)		x	x				Reviews
Dave et al. (2003)		x	x				Reviews
Yu and Hatzivassiloglou (2003)	x	x	x				News Items
Bai et al. (2004)		x					Reviews
Gamon (2004)		x	x				Customer feedback
Mullen and Collier (2004)		x					Reviews
Matsumoto et al. (2005)		x					Reviews
Read (2005)		x					Reviews
Riloff et al. (2006)		x	x				Reviews
Abbasi et al. (2008)		x	x				Reviews
Go et al. (2009)		x	x				Twitter
Prabowo and Thelwall (2009)	x	x	x				Reviews
Melville et al. (2009)	x						Reviews
Pak and Paroubek (2010)		x					Twitter
Barbosa and Feng (2010)	x		x				Twitter
Davidov et al. (2010)		x					Twitter
Kouloumpis et al. (2011)	x	x	x				Twitter
Taboada et al. (2011)	x						Reviews
Agarwal et al. (2011)	x	x	x				Twitter
Smeureanu and Bucur (2012)		x					Reviews
Wang and Manning (2012)		x					Reviews
Neri et al. (2012)		x					Facebook
Blamey et al. (2012)		x					Twitter
Kumar and Sebastian (2012)	x	x					Twitter
Ben Hamouda and El Akaichi (2013)		x					Facebook
Troussas et al. (2013)		x					Facebook
Tamilselvi and ParveenTaj (2013)		x	x				Twitter
Habernal et al. (2014)		x	x				Facebook
Ortigosa et al. (2014b)	x						Facebook
Basiri et al. (2014)	x	x			x		Reviews
da Silva et al. (2014)		x					Twitter
Fersini et al. (2014)		x					Reviews, Twitter
Yu and Wang (2015)		x					Twitter
Mohammad and Kiritchenko (2015)		x					Twitter
Our study	x	x	x	x	x	x	Facebook

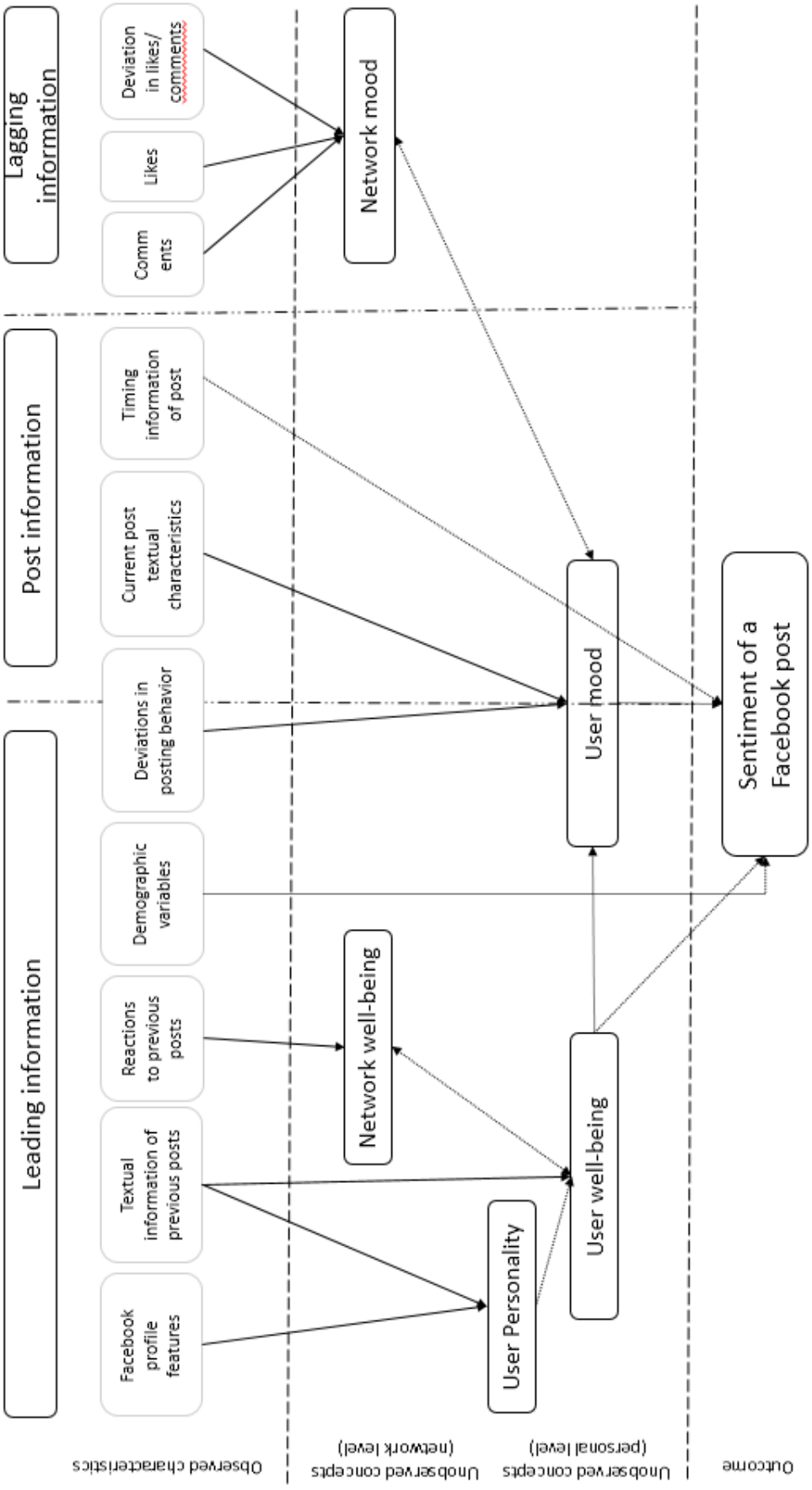


Figure 2.1 : Conceptual framework representing the literature review

2.1. Leading information

Leading Facebook information includes the complete history of a user's Facebook trail, including previous posts. We hypothesize that this information will improve sentiment classification prediction because several user characteristics can influence expressed sentiment. Settanni et al. (2015) show that textual indicators extracted from Facebook may be used to study subjective well-being, a result confirmed by Kramer (2010). This means that, by looking at previous posts of the same user and the valence of those posts, we can make an assumption about the subjective well-being of the user. Moreover, Diener (1998) states that personality is a major determinant of long term, subjective well-being. This is an important point, given that several researchers report that Facebook profile features (Kosinski et al, 2013; Ortigosa et al., 2014a) as well as text (Pennebaker et al., 2003) can accurately predict personality traits. By incorporating these Facebook profile features and previous textual features, we thus aim to incorporate the subjective well-being of a user as a predictor. As this is a long-term emotional state of a user, we believe subjective well-being can be informative of the sentiment expressed in Facebook posts.

While subjective well-being can add value, short-term changes (the 'mood' of a user) can affect sentiment of Facebook posts as well. Ortigosa et al. (2014b) state that behavior variations as shown on Facebook, can indicate changes in the user's mood. Smith and Petty (1996) report that positive or negative framing of a message could create more attention, especially in the case where the framing is unsuspected, as is the case with short-term changes from subjective well-being. We therefore argue that deviations from a user's average posting behavior can be informative of the sentiment of that post.

Comparable to a person's subjective well-being, we refer to network well-being as the overall emotional state of the network of the user. Network well-being and the focal user's well-being are connected by a phenomenon called emotional contagion (Christakis and Fowler, 2011; Helliwell and Putnam, 2004), which is defined as the tendency to automatically mimic other persons, and consequently to converge emotionally (Hatfield et al., 1994). This influence works in both ways. Network well-being can thus be informative about a user's well-being, and hence about the sentiment expressed in the user's Facebook posts. Quercia et al. (2012) already showed that community well-being can be predicted by using sentiment of community members' tweets. Since Facebook posts of the user's network were not available, we use the reactions to previous posts of the focal user to take into account part of network well-being that can be measured.

Finally, Schwartz et al. (2013) not only found differences in language usage across personalities, but also across gender and age. By incorporating these demographic variables and allowing for interaction effects, we assume that the textual features can bring even more added value to sentiment prediction.

Overall, the leading variables allow researchers to take into account heterogeneity among users with regards to word use, wellbeing, mood and demographics. The leading variables are discussed in detail in Section 3.4.2 and Table A2 in Appendix A. Fig. 1 shows the relationships described above in a visual way. The top panel shows the observed characteristics, the middle panel contains the unobserved, or latent, concepts, and the bottom panel represents the outcome. Solid lines represent the measurement model, while dotted lines are intended to show the structural model. For example, Facebook profile features are expected to, in part, measure user personality, while user personality influences user well-being and hence influences the sentiment of a Facebook post. For the sake of completeness, we also added the expected relationships for the focal post characteristics. The focal post's textual characteristics can be informative of the focal user's mood, while the timing variables are taken into account directly as control variables for post sentiment.

It is important to note that the concepts are introduced to provide plausible explanations of our findings about the relationship between the observed (top layer) characteristics and the outcome (bottom layer). Unfortunately our data do not allow us to model the concepts in the middle layer as our measurement model is incomplete. For example, there are more observed characteristics that make up the concept 'network well-being'. We do not have access to these additional characteristics and therefore it would be incorrect to make claims about that particular concept. The primary goal of our conceptual framework is to support our findings that focus on the top and bottom layer. Analysis of the middle layer is out of the scope of this research and requires additional data generated through questionnaires.

2.2. Lagging information

The lagging variables comprise information on the likes and comments of a post, as well as deviations from previous liking and commenting behavior on posts. Previous research has shown that more negatively oriented posts tend to attract more comments (Stieglitz and Dan-Xuan, 2012). This can be explained by negativity bias (Baumeister et al., 2001). Negativity bias is defined as the tendency to react stronger to very negative stimuli than to matched positive stimuli. In terms of engagement on posts, this means that people are more engaged with

negatively oriented posts, and are willing to put more effort in commenting on the post. On the other hand, we expect that the number of likes a post receives is positively correlated with positive sentiment, as a 'like' has an inherent positive dimension. Forest and Wood (2012) indeed indicate that positive posts receive more likes compared to negative ones. In the case of positively oriented posts, people might simply opt to like the status, instead of taking the effort to write a comment, thereby shifting responses from comments to likes (Stieglitz and Dang-Xuan, 2012). Next to the number of comments and likes, we also evaluate the valence of comments. Previous research on discussion forums and political weblogs revealed that negatively oriented posts are found to receive more negative comments, while positively oriented posts receive more positive comments (Dang-Xuan and Stieglitz, 2012; Huffaker, 2010).

In accordance with the concepts of user well-being and network well-being, we propose a similar concept 'network mood', comparable to individual mood. An individual's mood can influence network mood and vice versa (e.g., by posting status updates), by mechanisms such as emotional contagion and empathy. Network mood can thus be informative about a user's mood, and hence about the sentiment of posts from that user. Since we do not have network posts available, we measure part of the network mood by the likes and comments on statuses of the focal user. As mentioned in the previous paragraph, unsuspected framings create more attention and involvement (Smith and Petty, 1996). We therefore also add deviation variables, indicating if a post received more comments or likes than average for that specific user, to define network mood.

The earlier results and the theoretical framework mentioned above suggest that information on likes, comments, and deviations is very valuable to detect emotional valence of a status, and we thus hypothesize that lagging variables add predictive value to our model. The lagging variables are discussed in more detail in Section 3.4.3 and Table A3 in Appendix A. They are also shown in Fig. 1. In sum, to the best of our knowledge we believe that this study is the first to include auxiliary features in sentiment analysis models. Based on the conceptual framework outlined in our literature review, we hypothesize that those data will significantly increase the predictive performance of our models.

3. Methodology

3.1. Data

The data were gathered using a Facebook application in the period from June 1, 2014 to July 13, 2014. The application was created for a European soccer team and advertised several times on the soccer club's Facebook page. In order to stimulate usage of our application, the users could win a jersey of the soccer team. When launching the application, the Facebook user was presented with an authorization box, which specified the data that were being collected. It was clearly stated that the data were collected solely for academic purposes. Contact information was also provided in case there were any questions. Once the user authorized the application, it started to gather personal information (e.g., gender, age, location), information on engagement behavior (e.g., Facebook groups the user belongs to, Facebook page likes, Facebook events the user attended) and general Facebook behavior (e.g., uploaded photos, videos, links and posts) from the user using the Facebook API. In total, we were able to capture 100,227 posts. As the Facebook application focused on Flemish soccer fans, the main language of the status updates is Dutch. In subsequent analyses we discard all non-Dutch posts. The average number of words used in the statuses is 15, which is com-parable to the average number of words in tweets (Go et al., 2009). The main difference is in the maximum number of words, which goes up to 968 for our Facebook sample, while the maximum number of tweet characters is limited to 140. Detailed information about all the Facebook variables can be found in Section 3.4.2 and Appendix A.

3.2. Model description

In order to formally assess the additional value of auxiliary information over and above a focal post's content, we fit three models. The first model is the base model and reflects all the information of the present, where 'the present' refers to the time of posting (i.e., it contains the time and text variables of the post). The second model contains both information from the present and from the past by including the leading variables. The third model augments the second model with lagging variables, which adds a third time dimension to the model (i.e., the future). The choice of these three models is therefore motivated by practical reasons. We call model 1 the base model as our literature review pointed out that it reflects current practice. Model 2 has the prospect of improving predictive performance and can still be deployed in real-time. Finally, model 3 is expected to further improve performance but requires us to wait until the post has had enough time to gather comments and likes. Because model 2 can be used in real-time and model 3 cannot, it is practically relevant to determine the difference in performance between these two models. Formally, the models have the following forms:

Model 1: Status sentiment = f(focal post's content)

Model 2: Status sentiment = f(focal post's content)

+ f(leading variables)

Model 3: Status sentiment = f(focal post's content)

+ f(leading variables)

+ f(lagging variables)

The definition of Status sentiment is described in Section 3.3, while the different independent variables are described in Sections 3.4.1, 3.4.2 and 3.4.3. The functional form of the models is not specified as we use a data mining approach without pre-set functional form, which is explained more in detail in Section 3.5.

3.3. Dependent variable description

For the creation of our dependent variable, we follow the approach of distant supervision used by Read (2005), Go et al. (2009) and Pak and Paroubek (2010). This approach filters out emoticons from tweets, and uses these emoticons to represent positive and negative sentiment of a tweet. The emoticons thus serve as noisy emotion labels (Go et al., 2009). We list emoticons taken from Wikipedia (Wikipedia, 2015) and assign a positive or negative sentiment to the emoticon. Our sentiment variable is then constructed by comparing the emoticons in the post with our reference list. In case of ties (positive as well as negative emoticons occur), the label is assigned by majority voting.

This approach implies that only Facebook messages with emoticons can be used in the training phase, which leads to a total of 17,697 available status updates (of which 2078 were classified as negative and 15,619 as positive). In order to overcome class imbalance, we apply oversampling (Ballings et al., 2015).

To test the accuracy of using emoticons for sentiment detection, a random subset of 2000 status updates was manually labeled by two annotators. The inter-annotater agreement (also called Fleiss' j (Landis and Koch, 1977)) between the three labels (label obtained by emoticons, annotator 1 and annotator 2) is 0.74. This score can be defined as substantial (Landis and Koch, 1977), which indicates that emoticons can indeed be used as sentiment labels.

3.4. Independent variable description

Different categories of variables were used in this study. As discussed above, the nature of the variables constitutes a major contribution of this paper, and hence we will further elaborate on

the variables included. These can be divided into three categories: a focal post's variables, auxiliary leading variables and auxiliary lagging variables. A summary of all the variables can be found in Appendix A.

3.4.1. Focal post's variables

First, we extracted time-related variables of the post. These variables are the time, day and month of posting and a dummy variable to indicate whether the post occurred in a weekend. We include these variables as control variables (de Vries et al, 2012).

Second, in order to perform the sentiment classification task, we need to process the textual information so that it can serve as input to the model. As described before, there exist a variety of text features that can be taken into account. We include as much features as possible in our predictive models, in order to have a powerful base model to test our augmented models against.

First of all, we include lexicon-based features. These features are calculated using a (Dutch) sentiment lexicon (CLiPS, 2014). This lexicon gives a positive/negative weight to each word, as well as a subjectivity score. We then calculate the positive polarity, negative polarity, overall polarity and subjectivity for each status update by simply summing the polarity and subjectivity scores of each word in the status update. If negation words occur next to polarity words, we change the orientation of the polarity scores. These scores per status update are input features for the prediction model. Next, we use syntactic features. This includes the number of punctuations, exclamation marks, question marks, capital letters, characters and words. It also includes part-of-speech. Finally, we also create lexical features. We only include unigram features, as past research gives no conclusive evidence for the added value of higher order n-gram features (Dave et al., 2003; Go et al., 2009; Matsumoto et al., 2005; Pak and Paroubek, 2010; Pang et al., 2002). In order to create the unigram, we follow the approach by Coussement and Van den Poel (2008), Pak and Paroubek (2010), Cao et al. (2011) and D'Haen et al. (2016). In a first step, all special characters, emoticons and punctuation are removed. A tokenization is performed by splitting each status in distinct words using spaces as separators. Next, stopwords such as 'the' or 'a' are removed since these words are frequently used and hold little or no content information (Frakes and Baeza-Yates, 1992). Abbreviations are replaced using a dictionary and a spelling check is conducted in order to cope with the noisy nature of social media data. Indeed, users often use their cell phones to post status updates which leads to a higher frequency of misspellings and slang (Go et al., 2009; Ortigosa et al., 2014b). The next step is lemmatization, followed by synonym replacement in order to further reduce the vector space. As a final step, stemming is applied. With stemming, a word is stripped to the basic form (i.e.,

suffixes and prefixes are removed) (D'Haen et al., 2016; Kraaij and Pohlmann, 1994; Porter, 1980). This process results in a basic unigram (also called bag-of-words or document-term matrix). The unigrams obtained by the procedure described above are still very sparse. Therefore, we apply a feature reduction technique that reduces the number of features for input to the classification algorithm. We chose to work with Latent Semantic Indexing (LSI). This method is proposed by Deerwester et al. (1990) and reduces the original matrix in dimension by its first k principal component directions (Deerwester et al., 1990).

3.4.2. Leading variables

Leading variables can be subdivided into five groups, as outlined in Section 2. Facebook profile features contain engagement behavior (e.g., number of Facebook events attended) and general Facebook behavior (e.g., number of photos, videos). Age and gender are included as demographic variables. Previous post information will control for the user's and network well-being. This information includes average measures, e.g. average polarity of posts and average number of likes on previous posts. Deviations from previous post information can be informative about user mood. We use the following equation to calculate these deviation variables:

$$\Delta X_{i,T} = X_{i,T} - \bar{X}_{i,1\rightarrow t} \tag{2.1}$$

where X denotes the specific variables, i represents a user and T indicates the time of posting. We thus calculate for every post the deviation between the post's feature score and the average feature score for the user that posted. Example variables are the deviation in the number of words and the deviation in the number of positive and negative words of the post. A complete list can be found in Table A2 in Appendix A.

3.4.3. Lagging variables

Lagging variables can only be observed after the content was posted, which are, in the case of Facebook, likes and comments. We thus include the number of likes, the number of comments, the number of likes on comments and textual information from comments (e.g., the number of positive or negative words in comments, the number of words in comments) into our predictive model. Further-more, as for the leading variables, we calculate deviations from the normal liking or commenting behavior on posts of the focal user. This includes for example the deviation in the number of comments and the deviation in the number of likes. In order to calculate the lagging variables, we allow each post to gather likes and comments for seven days. We chose this particular time frame for three reasons. First, this limitation increases the

practical feasibility of our solutions as sentiment analysis is most valuable within a short time frame. Second, as such we give each post equal time to gather likes and comments. Third, as Fig. 2.2 shows, more than 99% of all comments are gathered during the first week. A complete list of all lagging variables can be found in Table A3 in Appendix A.

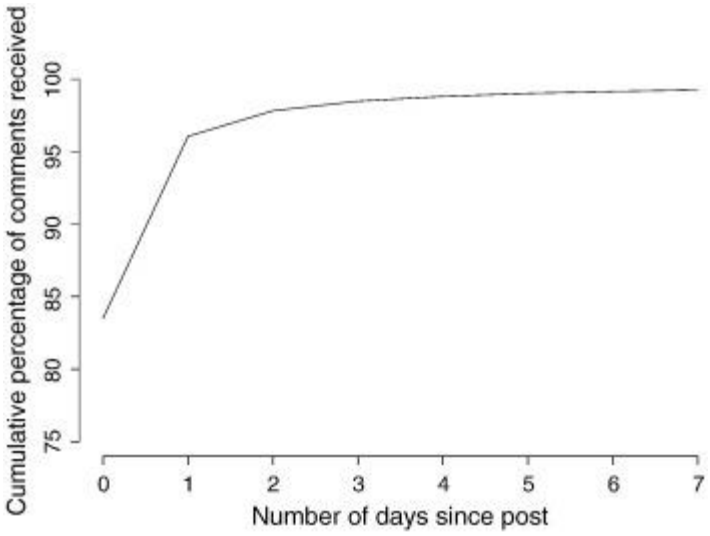


Figure 2.2: Cumulative collected % of comments per day

3.5. Predictive techniques

We use the Support Vector Machines (SVM) and Random Forest classification algorithms to perform our sentiment analysis. SVM has been used extensively in sentiment analysis and generally outperforms other methods such as Naïve Bayes, Maximum Entropy and logistic regression (Fernandez-Delgado et al., 2014). Although Random Forest classification has not been frequently used in sentiment analysis, it has recently been shown to be the best allround classification technique in many other domains (Fernandez-Delgado et al., 2014). Using both algorithms allows to use a well-established technique in sentiment analysis on the one hand, while on the other hand we can assess whether the Random Forest classification algorithm adds value in sentiment analysis.

3.5.1. Support Vector Machines

An important parameter in SVM is the kernel function (Bast et al., 2015). We use a radial basis (RBF) kernel, because this allows for non-linear relationships and requires the choice of only one hyperparameter γ, the width of the Gaussian (Ben-Hur and Weston, 2010). We thus have, combined with the SVM penalty parameter C, two parameters to choose. The choice of these parameters cannot be determined in advance. Hence, we follow the recommendation to test different values of C, ($C = [2^{-5}, 2^{-4}, \ldots, 215]$) and γ, ($\gamma = [2^{-15}, 2^{-14}, \ldots, 23]$) (Hsu et

al., 2003). We use the svm function of the e1071 R package (Meyer et al., 2014) to implement SVM.

3.5.2. Random Forest

The Random Forest classification algorithm grows a committee of classification trees and averages over all tree predictions (Breiman, 2001). By doing so, it can overcome the limited robustness and suboptimal performance of individual trees (Dudoit et al., 2002). Applying Random Forest has multiple advantages. It does not overfit (Breiman, 2001). Furthermore, it is easy to use in that variable importances are provided (Sandri and Zuccolotto, 2006) and only two parameters have to be set (Bogaert et al., 2016): the number of trees and the number of predictors to consider at each step in the tree. We set these parameters according to the guidelines of Breiman (2001) : the number of trees is set to 1000 and the number of predictors is defined as the square root of the total number of variables. Random Forest is implemented using the randomForest package in R provided by Liaw and Wiener (Liaw and Wiener, 2002).

3.6. Performance evaluation

Instead of classifying each post with a binary label {negative, positive}, we compute a score, representing the probability that a post is positive. For example, instead of saying that a post is positive, we would be able to say that the post is 70% likely to be positive, which is equivalent to saying that the post is 70% positive. Therefore, model performance is measured by the area under the receiver operating characteristic curve (AUC or AUROC). In case of scoring classifiers the AUC is a more adequate performance measure than, for example, accuracy as it does not rely on the cut-off values of the posterior probabilities (Ballings and Van den Poel, 2013). AUC is defined as follows:

$$AUC = \int_0^1 \frac{TP}{(TP+FN)} d \frac{FP}{(FP+TN)} = \int_0^1 \frac{TP}{P} d \frac{FP}{N}, \quad (2.2)$$

with TP: True Positives, FN: False Negatives, FP: False Positives, TN: True Negatives, P: Positives (positive sentiment), N: Negatives (negative sentiment). The values of the AUC range from 0.5 to 1. An AUC of 0.5 means that the model is not able to do better than a random selection, while a value of 1 indicates a perfect prediction (Ballings and Van den Poel, 2013).

3.7. Cross validation

We use five times twofold cross-validation (5x2 CV) (Alpaydin, 1999; Dietterich, 1998). This method randomly splits the sample into two partitions of equal size. The first partition serves as training set while using the second partition as test set and vice versa. This procedure is

repeated 5 times. Hence, a total of 10 performance measures per model will be obtained (Dietterich, 1998). We summarize these 10 performance measures with the median. To assess whether the AUCs of the different models are significantly different, we use the non-parametric Friedman test (Friedman, 1937) as suggested by Demšar (2006). The models are ranked, per fold separately, with the best model receiving the rank of 1, the second receiving the rank of 2 and the worst performing model receiving the rank of 3. In case of ties, the average rank is assigned. The Friedman statistic can then be defined as:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \quad (2.3)$$

where *N* is the number of folds, *k* is the number of models and *Rj* is the average rank of the *j*-th model over all folds.

3.8. Variable importance measures and Partial Dependence Plots

In order to interpret the relationships between independent variables and the sentiment classification, we will use the Random Forest models. The variable importances are assessed using the total decrease in node impurities from splitting on the variable, averaged over all trees in the Forest. The node impurity is measured by the Gini index *p(1 − p)*, and the decrease in node impurity is measured as follows:

$$\Delta(s,\tau) = p_\tau(1-p_\tau) - \left(\frac{|\tau_L|}{|\tau|} p_{\tau_L}(1-p_{\tau_L}) + \frac{|\tau_R|}{|\tau|} p_{\tau_R}(1-p_{\tau_R}) \right) \quad (2.4)$$

where *s* is short for a given split of a given variable and τ, τ_L, τ_R respectively stand for all the cases in the parent node, left child node and right child node. *p* is short for *p(y = 1)* with *y = {0, 1}* and thus denotes the probability that an observation is positive given that it is in that specific node. We denote cardinality by |•|. We use the importance function in the randomForest package in R (Liaw and Wiener, 2002). Remark that we take the median of the five times twofold cross-validated mean decrease in node impurity when we report importance measures.

Next to the most important variables, we are interested in the form of the relationship between predictors and the response. For this purpose, we use Partial Dependence Plots (Hastie et al., 2009). Partial Dependence Plots can be used to interpret any 'black box' model. Basically, the plots represent the relationship of one (or a subset) of the predictors with the response, taking into account the effect of all the other predictor variables. The Partial Dependence Plots are five times twofold cross-validated, using the interpretR R package (Ballings and Van den Poel, 2015).

4. Discussion of results

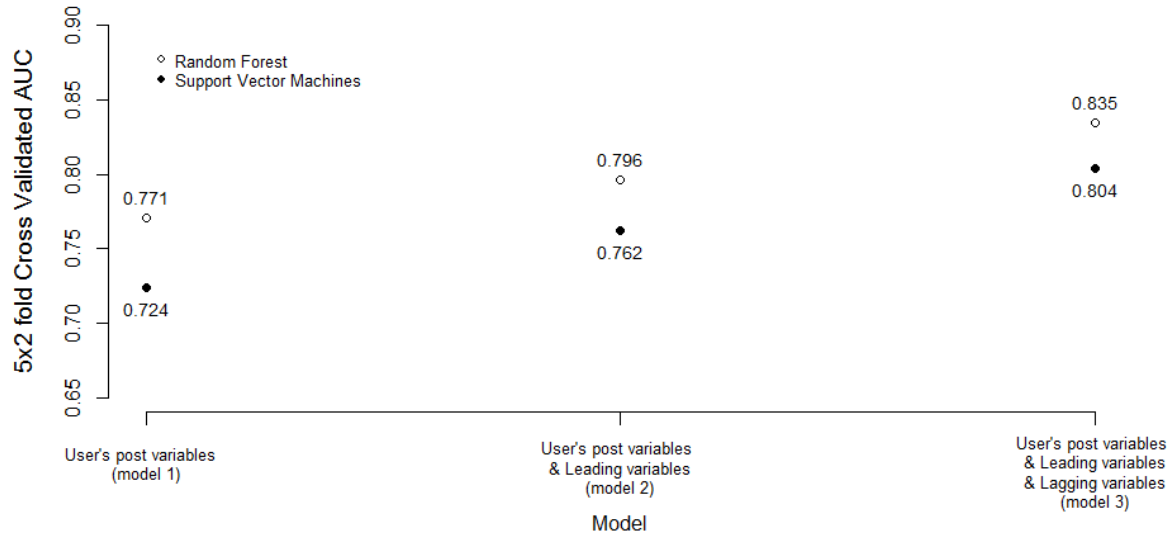


Figure 2.3: Result of the model in terms of AUC

As explained in Section 3.2, three models were built. The first model only considers the present information, the second model considers both present and past information and the third model considers present, past and future information. Fig. 2.3 shows the performance of the three models in terms of AUC, both for the Random Forest (solid line) and Support Vector Machine (dashed line) models. As the Random Forest algorithm creates better models across the board, all subsequent results will be discussed in terms of the Random Forest model. Remark that the reported AUCs are median values of the five times twofold cross-validation procedure.

The Friedman test indicates the presence of a significant difference in the analysis (χ_3^2 = 20, p < 0.01) Subsequently we made pairwise comparisons between the models and found that on each of the ten folds, the second model performs better that the first model, and the third model performs better than the second model. This means that model 2 is significantly better than model 1 (p=0) and that model 3 is significantly better than model 2 (p=0).

In sum, the AUCs show that leading and lagging variables add value to the user's post variables. In order to understand what drives these results, we analyzed the variable importances. The top 50 variable importances of the best, most comprehensive model (the third Random Forest model: user's post variables & leading variables & lagging variables) are shown

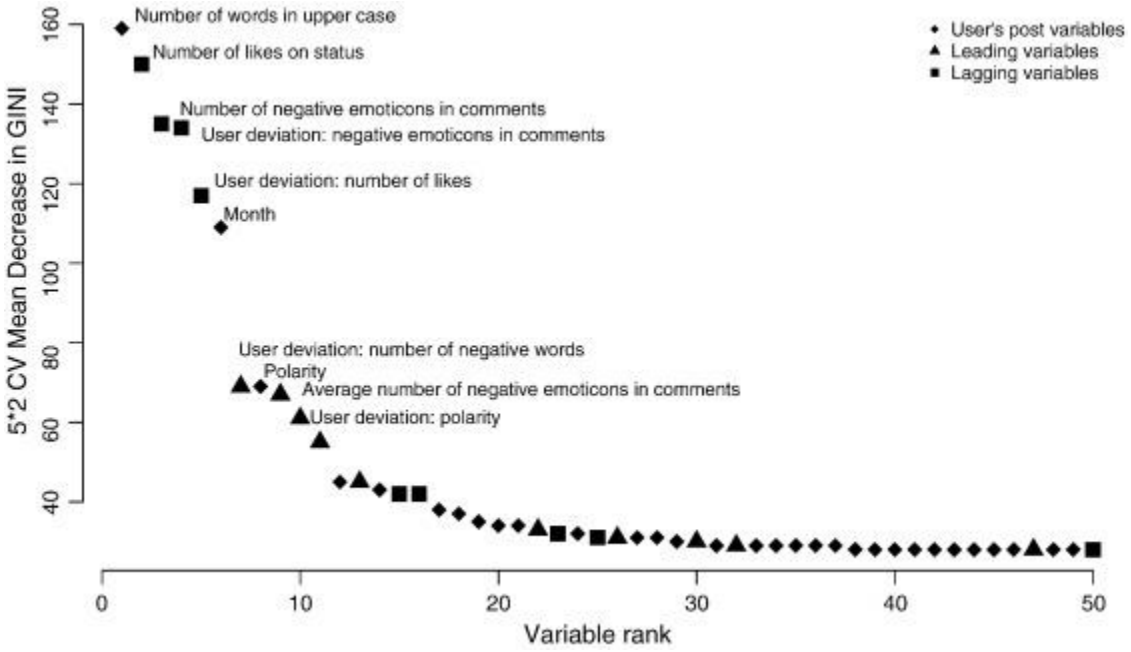


Figure 2.4: Variable importances of most complete model

in Fig. 2.4 and listed in Appendix B. In Fig. 2.4, the variables are sorted in descending order of (5x2 CV median) mean decrease in Gini, which means that the most important variables are ranked first. When looking at the graph, we see that the top 10 importances are mixed among the three components of the model; three variables originate from the user's post variables, three variables are leading variables and the remaining four variables contain lagging information. This again suggests that all data sources are complementary to each other. We will continue with a discussion of the top post, leading, and lagging variables, starting with the post variables. We use Partial Dependence Plots (PDPs) for this purpose. The PDPs depict the predicted probability of a positive post on the y-axis, and the different values of the predictor on the x-axis.

We see that the number of uppercase letters and the post polarity both have a positive relationship with positive sentiment, as depicted in Figs. 2.5a and 2.5b. For polarity, this was expected as it measures the positivity of a post based on the lexicon approach. Our research also suggests that the number of uppercase letters is strongly related to positive sentiment. Capital letters are used when users are more passionate about the post. They are often used as intensifiers of the message (Taboada et al., 2011). A look at the negative posts in our sample brings up a possible explanation for the positive direction of the intensifier. Negative posts on Facebook frequently convey low-arousal negative feelings (e.g., feeling sick, alone) instead of high-arousal feelings such as complaints or anger. This means that there is no need to use

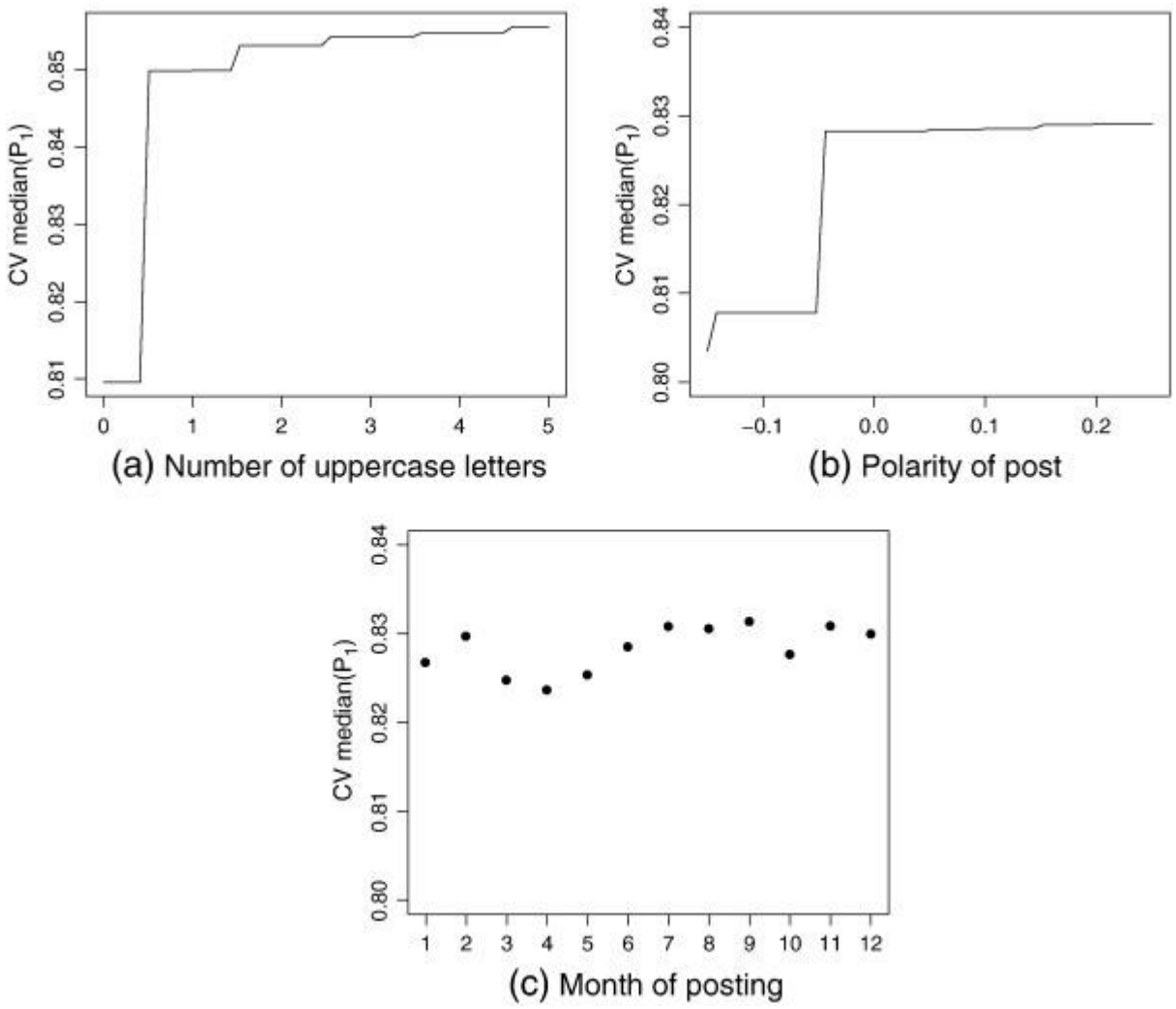


Figure 2.5: Partial Dependence Plots of post variables

intensifiers for these negative feelings, leaving intensifiers to be used mainly for positive posts. Although several papers include uppercase words or letters as features, none of the papers report the importance of the uppercase feature separately, making it impossible to compare our results. Finally, month of posting is an important predictor. The plot (Fig. 2.5c) does not show a clear pattern, except that spring months score a little bit lower than average. This can be caused by the relatively poor performance of the soccer team during this period. Indeed, a larger proportion of the posts is related to this soccer team compared to a completely random selection of posts. As such, this result is not immediately generalizable, but we show the importance of including timing variables as control variables in sentiment analysis. Finally, it is worth noting that Appendix B shows that 30 out of the top 50 variables are post variables.

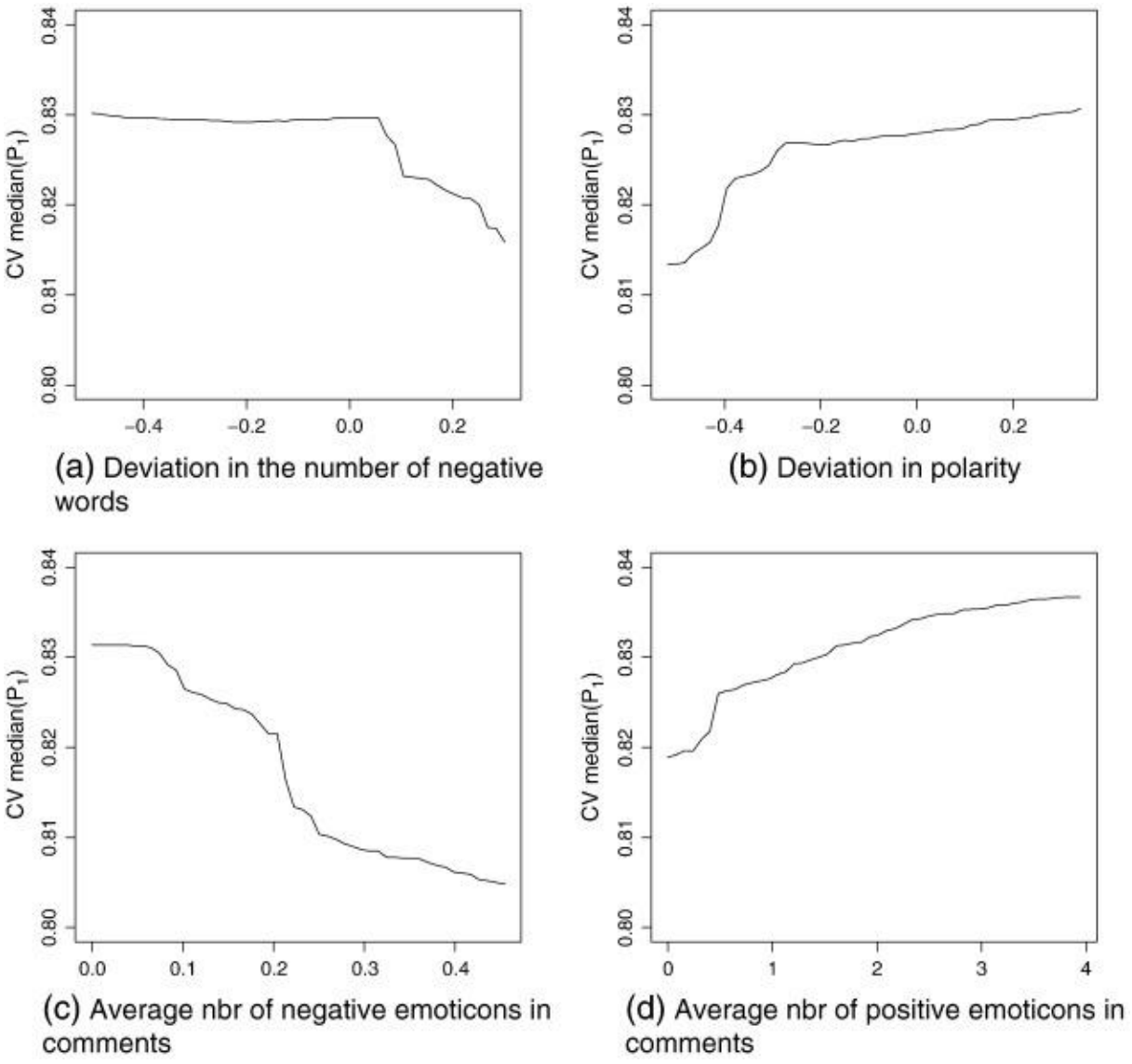


Figure 2.6: Partial Dependence Plots of main leading variables

Fig. 2.6 shows the Partial Dependence Plots for the leading variables. The deviation in the number of negative words and polarity are shown in the top row (Fig. 2.6a and 2.6b). A higher deviation in the number of negative words (i.e., more negative words are used than on average) leads to a higher probability of negative sentiment. A deviation in the number of negative words (i.e., more negative words are used than on average) leads to a higher probability of negative sentiment. A negative deviation in polarity leads to a higher probability of negative sentiment as well. This means that if the polarity of a post is more negative than the user's average post, the post will receive a more negative score. Fig. 2.6c and 6d shows the average number of negative/positive emoticons in comments (the average number of positive emoticons in comments is the eleventh most important variable). We see that a higher average number of positive/negative emoticons in comments on previous posts, indicates a higher probability of a positive/negative focal post. This supports our conceptual framework and indicates that well-

being can be predictive of sentiment. Furthermore, Fig. 2.6a and 2.6b indicate that also mood, as a temporal change of subjective well-being, can be informative. Indeed, Ortigosa et al. (2014a) state that behavior variations, such as deviations from the average polarity of posts shown in Fig. 6a and 6b, indicate changes in the user's mood. Finally, when looking at the top 50 most important variables, we see age as an important demographic variable, and the mean and standard deviation of the time between the focal user's page likes as important personality-related variables.

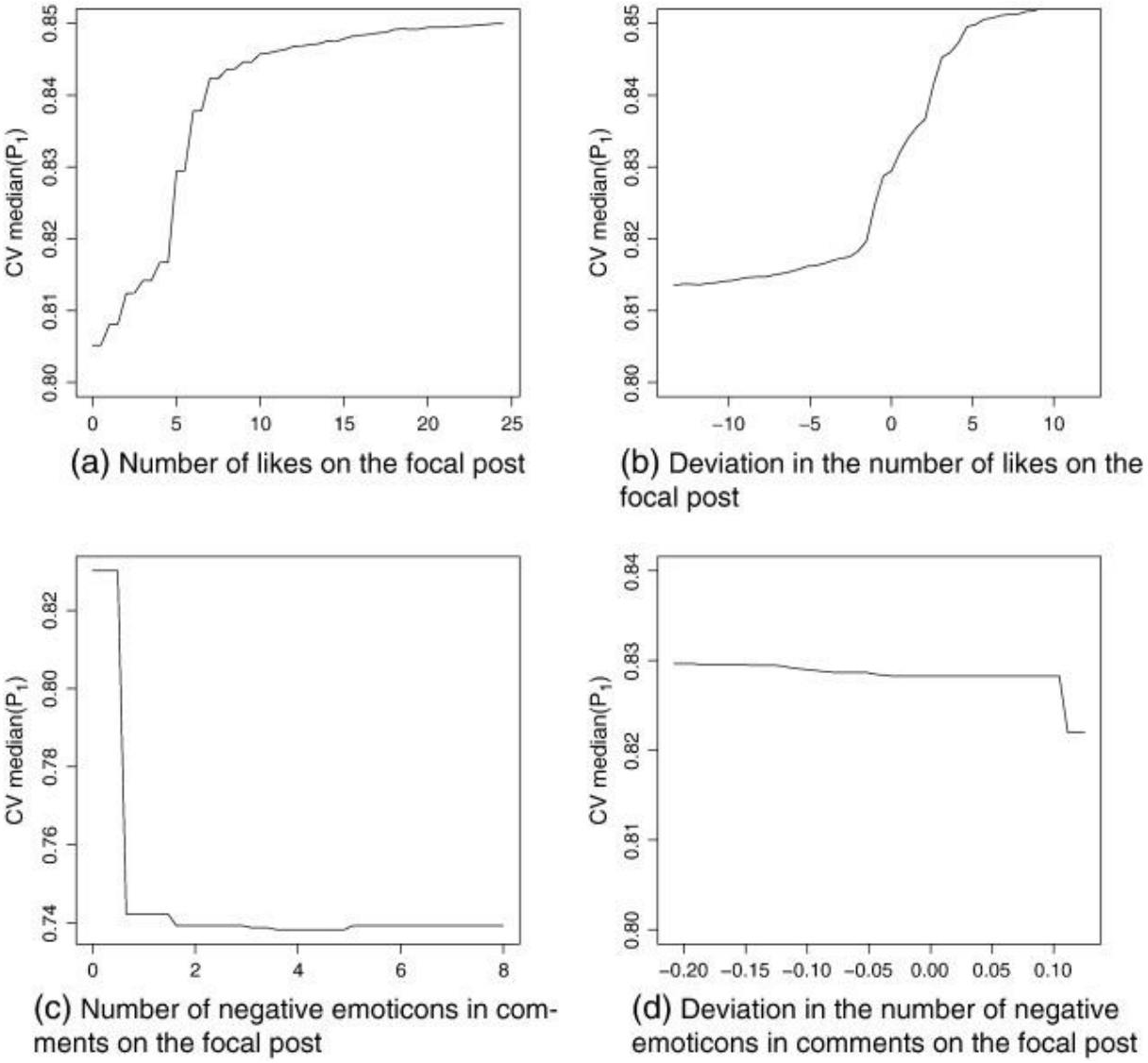


Figure 2.7: Partial Dependence Plots of main lagging variables

Finally, the top lagging variables are discussed. These are plotted in Fig. 2.7. While the number of likes (depicted in Fig. 2.7a) are very predictive, the number of comments do not seem that important (only fiftieth most important variable; not shown). The relationship of likes is as expected: the higher the number of likes, the higher the probability of positive sentiment. Fig.

7b shows the deviation in the number of likes compared to the average number of likes on posts by the same user. If the post receives less likes compared to an average post, the probability of positive sentiment declines. Fig. 2.7c and 2.7d show the number and deviation of negative emoticons in comments on the focal post. Both graphs shows that a higher number of negative emoticons, both in absolute figures and compared to the average number of the user, indicate a higher probability of negative sentiment. These results confirm the earlier findings of Stieglitz and Dang-Xuan (2012), and also support our conceptual framework relating to network mood, user mood and post sentiment. Stieglitz and Dang-Xuan (2012) also found a positive relationship between positive emoticons in comments and the positive sentiment of a post. We find this variable on a sixteenth place, with indeed a positive relationship (not shown), but of much smaller magnitude.

All previous results apply to a model trained and tested on posts with emoticons, which are used as noisy labels. These posts may be easier to predict than regular posts, because they express clear and strong emotions. Therefore, we manually labeled a random sample of 2000 posts without emoticons, and tested the model on these posts. The inter-annotator agreement (Fleiss' κ) for the statuses is 0.81, indicating that the task was well-defined (Landis and Koch, 1977). The annotators dis-agreed in 198 cases, which were subsequently revised and assigned a final sentiment label in order to include them in the analysis. For subsequent analysis, we dropped neutral statuses (259 cases) (Dave et al., 2003; Go et al., 2009; Pang et al., 2002). In that way, we can apply our model to the new statuses, which are used as new test samples for each of the folds. Results showed that model 1 achieved a median AUC of 0.751, model 2 a median AUC of 0.775 and model 3 a median AUC of 0.812. We can conclude that (1) the focal post's variables show significantly lower performance compared to models using statuses with emoticons, probably because emotions are expressed less clearly and (2) there is an effect of both leading and lagging variables. The effects in terms of extra predictive power are very similar to the case of statuses with emoticons. In summary, the results for posts with and without emoticons are very similar and consistent in terms of the added value of leading and lagging information.

5. Conclusion and practical recommendations

Initially, sentiment analysis was performed mainly on review data. Recently, because of their abundance, social media data have become the main focus in the field. Despite this change in focus, our literature review shows that researchers have not yet explored the additional wealth of information that is available through social media data. Therefore, in this study we set out to

(1) study the added value of leading and lagging variables for sentiment analysis, (2) determine the top predictors, (3) and explore the relationships of the top predictors with the sentiment of a post. We devised a conceptual framework to support our results.

The results clearly indicate that leading and lagging variables add predictive value to established sentiment analysis models. In other words, past and future information does add value over present information. The magnitude of the differences in model performance and the consistency of these differences over all folds suggest that the results are relevant. Given that Facebook messages are informal and therefore often contain slang, irony or multi-lingual words (Ortigosa et al., 2014b), sentiment analysis is difficult based solely on text. We showed that leading and lagging variables can help to predict sentiment in this challenging environment, and our conceptual framework helped in explaining why these variables matter.

The most important predictors of the most complete model were a mix of post variables (e.g., number of uppercase letters), leading variables (e.g., average number of negative comments on posts in the past) and lagging variables (e.g., number of likes) indicating that all three model components add to the predictive value of our model. We can draw several conclusions from these findings.

First, we can see that word use and time of posting are important. The number of uppercase letters is the most important predictor, followed by month of posting and the use of negative and positive words (polarity) as the sixth and eighth most important factors, respectively. Moreover, we see that a deviation in polarity is important, indicating a mood change from the general subjective well-being of the user, thereby supporting our predictions based on the conceptual framework. Finally, in total 30 of the 50 most important variables are related directly to the post's content and time of posting.

Second, it becomes clear that reactions on status updates contain relevant information, as 6 out of the 10 most important predictors stem from likes and comments related variables. A higher number of likes indicates a more positive post, while negative emoticons in the comments (on the current post, on previous posts, and deviations from previous posts) indicate negative posts. It thus seems that there is additional information in the variables that measure network well-being and mood. This also confirms previous findings from Stieglitz and Dang-Xuan (2012).

Third, we can conclude that general Facebook variables and demographics seem less important. Age is the thirteenth most important variable, while only two Facebook-related

variables show up in the top 50 (the average and standard deviation in page liking behavior of the user). Page liking behavior has already been shown to be predictive of, among others, happiness and personality traits (Kosinki et al., 2013), and thus user well-being, which makes this result plausible. The implication is that one could save the burden to gather the immense amount of data from Facebook, as the majority of the variables have only limited importance. Based on our results, we thus argue that age, page liking behavior and of course posts of the user are the most important Facebook variables to identify.

Finally, we would like to make a general remark on the importance of variables. We see that negative variables receive more attention from the algorithm than positive variables, or that deviations in the negative direction have a bigger influence. This can be linked to the lower number of negative posts in our sample and on Facebook in general (Lin Qiu, 2012; Newman et al., 2011). As the majority of the posts is positive, clues about negative sentiment turn out to be, in general, more useful to the algorithm. Therefore, we conclude that in a setting where the ratio of positive versus negative posts is high, features that indicate negativity can be more helpful to predict overall sentiment.

Academics, companies and public parties are interested in large scale sentiment analysis, which yields a wide range of applications. Companies can perform sentiment analysis to analyze customer satisfaction (Go et al., 2009), to increase ad-targeting efforts or to track public opinion about the company. Teachers can use sentiment analysis to support personalized e-learning (Ortigosa et al., 2014b). Academics measure general public mood and track changes over time. Political parties employ social media to track public sentiment and adjust their campaign towards regions or topics that suffer from negative emotions. Finally, broadcasters and media can analyze tweets to predict election outcomes (Tumasjan et al., 2010).

Established approaches to sentiment analysis described above include only present information. We propose to include all information from the past, which includes previous posts from the same user, in any sentiment analysis model. Indeed, even real-time applications can include leading information and benefit from the extra predictive value. Live television, for example, can analyze reactions on the Facebook or Twitter page in real-time, thereby including leading information. Another example may be news channels that analyze tweets real-time to predict elections (e.g., on the day of election), thereby using leading information. This could enable a more accurate prediction and better reputation of the news channel. On the other hand, real-time applications cannot benefit from lagging variables. However, other applications can take advantage of these lagging variables. For example, a company can allow for a small lag in

the measurement of customer satisfaction. This study used a lag of 7 days, but as Fig. 2 shows, more than 95% of all comments are gathered after only one day. The time frame for creating the lagging variables can thus be shortened, without losing much of the information. Finally, one can use the present and past information in a first round to quickly get an idea of the sentiment, and refine these early findings with lagging information in a second round. One possible application is a marketing campaign for a new product. First, the company can perform sentiment analysis to assess global sentiment concerning the product. In this way, the broad outlines of the marketing campaign can be adjusted if necessary. Second, more fine-grained sentiment analysis, including lagging variables, can be performed that allows to fine-tune the campaign. In sum, we feel that our proposed approach is a promising path for many sentiment analysis applications.

6. Limitations and future research

Sentiment analysis can be applied to a wide range of sources. Our research shows that leading and lagging information can be very valuable in the context of sentiment analysis onfFacebook posts. It remains unclear whether a similar approach can work for other media such as Twitter and review data, but we argue that the central idea is generalizable. Indeed, Twitter also includes leading information such as a concise user profile and previous tweets, while retweets and favorites can be seen as lagging information embedded in Twitter (e.g., Stieglitz and Dang-Xuan, 2013). An interesting avenue for further research would thus be to extend the application to other social media platforms.

Although our study extends the use of data that is available in social media to predict sentiment, and includes emotional contagion to some extent, we did not include complete network information in the analysis. Network effects are, to the best of our knowledge, not yet discussed in the area of sentiment analysis. However, there is a growing amount of research on social networks reporting the importance of network effects on a wide range of behaviors (e.g., Bakshy et al., 2012). As the main drivers of these effects are homophily and social influence (Hartmann et al., 2008), it can be expected that a user's emotions are related to the emotions of a user's network. Further research could try to incorporate network data and improve our results.

The third direction for future research is to use a more theoretical angle to approach the problem, while our primary goal was to look at the added value of leading and lagging variables taking a data mining approach. With the current results, it can be interesting to take a look at

the underlying constructs of (individual and network) well-being, mood and personality, and incorporate these constructs rather than all Facebook variables separately (e.g., by using a questionnaire). In this study we use latent constructs to provide plausible explanations of our findings about the relationship between the observed characteristics and the out-come variable, sentiment. As mentioned in the literature review, our data do not allow us to model the latent constructs as our measurement model is incomplete. We work with observed data and retrofitted latent constructs on these variables. Future research could start from latent constructs and make sure appropriate variables are included to fully measure each construct, which would allow for a formal measurement model. A logical approach would be to use data generated through surveys and use appropriate measurement scales. Because this study uses observed data we are unable to sort this out. Nevertheless, the unobserved concepts allow us to strengthen the theoretical underpinnings of our study, and facilitate the discussion of our results. We also feel that our conceptual model is a good basis for future theoretical and empirical research.

The fourth limitation is selection effects. It might be possible that the users whose information was obtained by using the application may be different from users that did not use the application. The Facebook application was developed for a European soccer team, which means that the users of the application are interested in soccer. This can also have its repercussions on the posts that are analyzed (i.e., they may be more soccer-oriented than the average Facebook post). In our opinion, this does not impose serious repercussions on the obtained results. In the case the posts are more biased towards one domain (e.g., soccer), it is likely that the text variables become more predictive because posts are more related and that sentiment is easier to predict (Ortigosa et al., 2014b). In this context, we were able to substantially improve our predictions by adding leading and lagging information. In case the domain is less bounded, it is likely that leading and lagging information can have even more predictive value.

The fifth limitation of this study is the limited number of values that some of the variables can have. Facebook limits the number of occurrences of a variable (e.g., the likes of a user) to the 25 most recent entries. This issue is most important for frequency variables that are included as part of the user profile information (which is part of the leading information). In order to deal with this limitation, we calculated frequency within a specific period of time. The length of this time window per variable is determined as to no user in our database reaches the maximum number of 25 entries.

As a final remark we want to say that although this study has some shortcomings, it is the first sentiment analysis study using such a variety of data. We feel that this is a valuable contribution to literature.

7. References

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums, ACM Transactions on Information Systems 26 (3) 12:1–12:34.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., 2011. Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 30–38.

Alpaydin, E., 1999. Combined 5 times 2 cv F Test for comparing supervised classification learning algorithms, Neural Computation 11 (8) 1885–1892.

Bai, X., Padman, R., Airoldi, E., 2004. Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket, Technical report, Carnegie Mellon University, School of Computer Science, Technical Report CMU-IS-RI-04-127.

Bakshy, E., Eckles, D., Yan, R., Rosenn, I., 2012. Social Influence in Social Advertising: Evidence from Field Experiments, Proceedings of the 13th ACM Conference on Electronic Commerce, ACM, New York, NY, USA, pp. 146–161.

Ballings, M., Van den Poel, D., 2013. Kernel factory: an ensemble of kernel machines, Expert Systems with Applications 40 (8) 2904–2913.

Ballings, M., Van den Poel, D., 2015. interpretR: Binary Classifier and Regression Model Interpretation Functions, June 2015.

Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R. 2015. Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications 42 (20) 7046–7056.

Barbosa, L., Feng, J., 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 36–44.

Basiri, M.E., Ghasem-Aghaee, N., Naghsh-Nilchi, A.R., 2014. Exploiting reviewers comment histories for sentiment analysis, Journal of Information Science 40 (3) 313–328.

Bast, E., Kuzey, C., Delen, D., 2015. Analyzing initial public offerings' short-term performance using decision trees and SVMs, Decision Support Systems 73 15–27.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D., 2010. Bad is stronger than good, Review of General Psychology 5 (4) 323–370.

Ben Hamouda, S., El Akaichi, J., 2013. Social networks text mining for sentiment classification: the case of Facebook statuses updates in the Arabic Spring Era, International Journal of Application or Innovation in Engineering and Management, 2 (5) 470–478.

Ben-Hur, A., Weston, J., 2010. A User Guide to Support Vector Machines, in: O. Carugo, F. Eisenhaber (Eds.), Data Mining Techniques for the Life Sciences, 609, Humana Press, Totowa, NJ, pp. 223–239.

Blamey, B., Crick, T., Oatley, G., 2012. R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora, in: M. Bramer, M. Petridis (Eds.), Research and Development in Intelligent Systems XXIX, Springer London. pp. 207–212.

Bogaert, M., Ballings, M., Van den Poel, D., 2016. The added value of Facebook friends data in event attendance prediction, Decision Support Systems 82 26–34.

Bollen, J., Pepe, A., Mao, H., 2009: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, . In L. Adamic, R. Baeza-Yates, and S. Counts (eds.), Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Palo Alto, CA: AAAI Press, pp. 450-453

Breiman, L., 2001. Random Forests, Machine Learning 45 (1) 5–32.
Cao, Q., Duan, W., Gan, Q., 2011. Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach, Decision Support Systems 50 (2) 511–521.
Chaovalit, P., Zhou, L., 2005. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005. HICSS '05, pp. 112c-112c.
Christakis, N.A., Fowler, J.H., 2011. Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives - How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do, reprint edition ed., Back Bay Books, New York, NY
CLiPS,2014. Pattern - Web mining module for Python, Antwerp University, URL https://github.com/clips/pattern.
Coussement, K., Van den Poel, D., 2008. Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Sytems with Applications 34 (1) 313–327.
da Silva, N.F.F., Hruschka, E.R., Hruschka, E.R., Jr, 2014. Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 170–179.
Dang-Xuan, L., Stieglitz, S., 2012. Impact and Diffusion of Sentiment in Political Communication An Empirical Analysis of Political Weblogs, Sixth International AAAI Conference on Weblogs and Social Media, pp. 1–4.
Dave, K., Lawrence, S., Pennock, D.M., 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of the 12th International Conference on World Wide Web, ACM, New York, NY, USA, pp. 519–528.
Davidov, D., Tsur, O., Rappoport, A., 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 241–249.
de Vries, L., Gensler, S., Leeflang, P.S.H., 2012. Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing, Journal of Interactive Marketing 26 (2) 83–91.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K, Harshman, T.K., 1990. Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) 391–407.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 1–30.
D'Haen, J., Van den Poel, D., Thorleuchter, D., Benoit, D.F., 2016. Integrating expert knowledge and multilingual web crawling data in a lead qualification system, Decision Support Systems 82 69–78.
Diener, E., 1998. Subjective Well-Being and Personality, in: D.F. Barone, M. Hersen, V.B.V. Hasselt (Eds.), Advanced Personality. The Plenum Series in Social/Clinical Psychology, Springer, US, pp. 311–334.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) 1895–1923.
Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (457) 77–87.
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15 (1) 3133-3181.
Fersini, E., Messina, E., Pozzi, F.A., 2014. Sentiment analysis: Bayesian ensemble learning, Decision Support Systems 68 26–38.
Forest, A.L., Wood, J.V., 2012. When Social Networking Is Not Working: Individuals With Low Self-Esteem Recognize but Do Not Reap the Benefits of Self-Disclosure on Facebook. Psychol Sci 23, 295–302.

Frakes, W., Baeza-Yates, R., 1992. Information Retrieval: Data Structures and Algorithms, Prentice Hall PTR.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (200) 675–701.

Gamon, M., 2004. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis, Proceedings of the 20th International Conference on Computational Linguistics, No. 841, Association for Computational Linguistics. Vol. No. 841 of COLING 04, Stroudsburg, PA, USA, pp. 1–7.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision, Technical report, CS224N Project Report, Stanford

Habernal, I., Ptek, T., Steinberger, J., 2014. Supervised sentiment analysis in Czech social media, Information Processing & Management 50 (5) 693–707.

Hartmann, W.R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., Tucker, C., 2008. Modeling social interactions: identification, empirical methods and policy implications, Marketing Letters 19 (3-4) 287–304.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics Springer New York, New York, NY

Hatfield, E., Cacioppo, J.T., Rapson, R.L., 1994. Emotional contagion, Studies in emotion and social interaction. vii. Editions de la Maison des Sciences de l'Homme, Paris, France

Hatzivassiloglou, V., Wiebe, J.M., 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity, Proceedings of the 18th Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 299–305.

Helliwell, J.F., Putnam, R.D., 2004. The social context of well-being, Philosophical Transactions of the Royal Society B: Biological Sciences 359 (1449) 1435–1446.

Hsu, C.-W. , Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification, Tech. rep., Department of Computer Science, National Taiwan University.

Huffaker, D., 2010. Dimensions of Leadership and Social Influence in Online Communities, Human Communication Research 36 (4) 593–617.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media, Business Horizons 53 (1) 59–68.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., Graepel, T., 2013. Manifestations of user personality in website choice and behaviour on online social networks, Machine Learning 95 (3) 357–380.

Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior, Proceedings of the National Academy of Sciences 110 (15) 5802–5805.

Kouloumpis, E., Wilson, T., Moore, J., 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011. pp. 538–541.

Kraaij, W., Pohlmann, R., 1994. Porters stemming algorithm for Dutch, Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie, pp. 167–180.

Kramer, A.D., 2010. An Unobtrusive Behavioral Model of "Gross National Happiness", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, pp. 287–290.

Kumar, A., Sebastian, T.M., 2012. Sentiment analysis on Twitter, IJCSI International Journal of Computer Science Issues 9 (4(3)) 372–378.

Landis, J.R., Koch, G.G., 2010. The measurement of observer agreement for categorical N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems 48 (2) 354–368.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest, R News 2 (3) 18–22.

Lin Qiu, H.L., 2012. Putting their best foot forward: emotional disclosure on Facebook, Cyberpsychology, behavior and social networking 15 (10) 569–572.

Liu, B., 2012. Synthesis Lectures on Human Language Technologies, Sentiment Analysis and Opinion Mining 5, Morgan & Claypool Publishers., pp. 1–167.

Martínez-Cámara, E., Martín-Valdivia, M.T., Ure na López, L.A., Montejo-Ráez, A., 2014. Sentiment analysis in Twitter, Natural Language Engineering 20(1) 1–28.

Matsumoto, S., Takamura, H., Okumura, M., 2005. Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees, in: T.B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. No. 3518 in Lecture Notes in Computer Science, Springer, Berlin Heidelberg, pp. 301–311.

McInnes, B., 2009. Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text Using the Umls and Metamap, University of Minnesota, Minneapolis, MN, USA. (Ph.D. thesis)

Melville, P., Gryc, W., Lawrence, R.D., 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2009, pp. 1275–1284.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2014. C++-code), C.-C. C. l., C++-code), C.-C. L. l., Sep. 2014.

Mohammad, S.M., Kiritchenko, M., 2015. Using hashtags to capture fine emotion categories from Tweets, Computational Intelligence 31 (2) 301–326.

Mullen, T., Collier, N., 2004. Sentiment analysis using support vector machines with diverse information sources, Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 412–418.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T., 2012. Sentiment Analysis on Social Media, Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, Washington, DC, USA, pp. 919–926.

Newman, M.W., Lauterbach, D., Munson, S.A., Resnick, P., Morris, M.E., 2011. It's Not That I Don'T Have Problems, I'M Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health, Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, pp. 341–350.

Ortigosa, A., Carro, R.M., Quiroga, J.I., 2014. Predicting user personality by mining social interactions in Facebook, Journal of Computer and System Sciences 80 (1) 57–71.

Ortigosa, A., Martin, J.M., Carro, R.M., 2014. Sentiment analysis in Facebook and its application to e-learning, Computers in Human Behavior 31 527–541.

Pak, A., Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of LREC,2010, 2010. pp. 1320–1326.

Pang, N., Lee, L., 2008. Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval 2. No. 2 in 2., Now Publishers Inc., pp. 1–135.

Pang, N., Lee, L., Vaithyanathan, S., 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 79–86.

Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G., 2003. Psychological aspects of natural language use: our words, our selves, Annual Review of Psychology 54 (1) 547–577.

Porter, M., 1980. An algorithm for suffix stripping, Program 14 (3) 130–137.

Prabowo, R., Thelwall, M. 2009. Sentiment analysis: a combined approach, Journal of Informetrics 3 (2) 143–157.

Quercia, D., Ellis, J., Capra, L., Crowcroft, J., 2012. Tracking "Gross Community Happiness" from Tweets, Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, pp. 965–968.

Read, J., 2005. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, Proceedings of the ACL Student Research Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48.

Riloff, E., Patwardhan, D., Wiebe, J., 2006. Feature Subsumption for Opinion Analysis, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for computational Linguistics, Stroudsburg, PA, USA, pp. 440–448.

Sandri, M., Zuccolotto, P., 2006. Variable Selection Using Random Forests, in: P.S. Zani, P.A. Cerioli, P.M. Riani, P.M. Vichi (Eds.), Data Analysis, Classification and the Forward Search. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg, pp. 263–270.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach, PLoS ONE 8 (9). e73791.

Settanni, M., Marengo, D., 2015. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts, Personality and Social Psychology, 1045.

Smeureanu, I., Bucur, C., 2012, Applying supervised opinion mining techniques on online user reviews, Informatica Economica 16 (2) 81–91.

Smith, S.M., Petty, R.E., 1996. Message framing and persuasion: A message processing analysis, Personality and Social Psychology Bulletin 22 (3) 257–268.

Stieglitz, S., Dang-Xuan, L., 2012. Impact and diffusion of sentiment in public communication on Facebook, ECIS 2012 Proceedings

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

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2) 267–307.

Tamilselvi, A., ParveenTaj, M., 2013. Sentiment analysis of micro blogs using opinion mining classification algorithm, International Journal of Science and Research (IJSR) 2 (10) 196–202.

Tan, S., Wang, Y., Cheng, C., 2008. Combining Learn-based and Lexicon-based Techniques for Sentiment Detection Without Using Labeled Examples, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, pp. 743–744.

Troussas, C., Virvou, M., Junshean Espinosa, K., Llaguno, K., Caro, J. ,2013. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning, 2013 Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–6.

Tumasjan, A., Springer,T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178–185.

Turney, P.D., 2002. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 417–424.

Wang, S., Manning, C.D., 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 90–94.

Wikipedia, List of emoticons, 2015, URL http://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=654618502

Yu, H., Hatzivassiloglou, V., 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the 2003

Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 129–136.

Yu, Y., Wang, X., 2015. World Cup 2014 in the Twitter world: a big data analysis of sentiments in US sports fans' tweets, Computers in Human Behavior 48, 392–400.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B. 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis, HP Laboratories

8. Appendix

Appendix A: Variable list

Table A1
Focal post's variables.

Variable name	Variable description (Category)
SVD concept 1 - 100	SVD concepts (Lexical)
Number_uppercase	Number of uppercase letters in post (Lexical)
Number_punct	Number of punctuations in post (Lexical)
Number_qm	Number of question marks in post (Lexical)
Number_em	Number of exclamation marks in post (Lexical)
Number_nbr	Number of numbers in post (Lexical)
Number_wow	Number of 'wow' (or similar like 'woooow') mentioned in post (Lexical)
Number_pf	Number of 'Pf' (or similar like 'Pffff') mentioned in post (Lexical)
Number_lol	Number of 'lol' mentioned in post (Lexical)
Number_characters	Number of characters in post (Lexical)
Number_words	Number of words in post (Lexical)
Number_pos_words	Number of positive words in post (Lexicon)
Number_neg_words	Number of negative words in post (Lexicon)
Positive_polarity	Sum of positive polarity scores for the post (Lexicon)
Negative_polarity	Sum of negative polarity scores for the post (Lexicon)
Polarity	Sum of polarity scores for the post (Lexicon)
Subjectivity	Sum of subjectivity scores for the post (Lexicon)
POS_noun	Number of nouns in post (Syntactic)
POS_verb	Number of verbs in post (Syntactic)
POS_adj	Number of adjectives in post (Syntactic)
Month	Month of post (Time)
Weekday	Day of week of post (1 to 7) (Time)
Weekend	Dummy indicating if post occurred during weekend (Time)
Time_of_day	Time of the day of post (Time)

Table A2
Leading variables.

Variable name	Variable description (category)
Previous post information	
Mean_neg_emo	Average number of negative comments received on previous posts
Mean_pos_emo	Average number of positive comments received on previous posts
Mean_likes_posts	Average number of likes received on previous posts
Mean_comm_posts	Average number of comments received on previous posts
Mean_comm_likes_user	Average number of comments received on previous posts, liked by the user
Total_nbr_likes	Total number of likes received on previous posts
Total_nbr_comments	Total number of comments received on previous posts
Mean_polarity	Mean polarity of previous posts
Mean_pos_words	Mean number of positive words in previous posts

Mean_neg_words	Mean number of negative words in previous posts
Mean_subjectivity	Mean subjectivity of previous posts
Mean_nbr_words	Mean number of words in previous posts
Deviation_polarity	Deviation in polarity of the focal status compared to previous posts
Deviation_pos_words	Deviation in number of positive words compared to previous posts
Deviation_neg_words	Deviation in number of negative words compared to previous posts
Deviation_subjectivity	Deviation in subjectivity of the focal status compared to previous posts
Deviation_nbr_words	Deviation in number of words in the focal status compared to previous posts
Total_nbr_posts	Total number of previous posts

General Facebook information

Age	Age of user (personal information)
Gender	Gender of user (personal information)
Relationship_single	Dummy indicating whether the person is in a relationship or not (personal info)
Heterosexual	Dummy indicating whether the person is heterosexual (personal information)
Account_age	Age of the Facebook account of the user (personal information)
Number_friends	Number of friends of the user (personal information)
Number_groups	Number of Facebook groups the user is member of (engagement behavior)
Number_likes	Number of Facebook pages the user has liked (engagement behavior)
Number_events	Number of Facebook events the user has attended (engagement behavior)
Number_interests	Number of interests as expressed on Facebook (engagement behavior)
Number_check-ins	Number of check-ins registered on Facebook (engagement behavior)
Number_cin_likes	Number of likes on check-ins (engagement behavior)
Number_cin-tags	Number of tags related to check-ins (engagement behavior)
Number_cin_comments	Number of comments related to check-ins (engagement behavior)
Number_photos	Number of photos (general FB behavior)
Number_videos	Number of videos (general FB behavior)
Number_links	Number of links (general FB behavior)
Number posts	Number of posts (general FB behavior)
Number_comm_photos	Number of comments received on photos (general FB behavior)
Number_comm_videos	Number of comments received on videos (general FB behavior)
Number_comm_links	Number of comments received on links (general FB behavior)
Number_likes_photos	Number of likes received on photos (general FB behavior)
Number_likes_videos	Number of likes received on videos (general FB behavior)
Number_likes_links	Number of likes received on links (general FB behavior)
Recency_comment	Recency of comments received from other users (general FB behavior)
Recency_likes	Recency of likes received from other users (general FB behavior)
Recency_photo	Recency of last photo at time of post posting (general FB behavior)
Recency_video	Recency of last video at time of post posting (general FB behavior)
Recency_link	Recency of last link at time of post posting (general FB behavior)
Recency_check-in	Recency of last check-in at time of post posting (general FB behavior)
Recency_like	Recency of last page like at time of post posting (general FB behavior)
Recency_post	Recency of last post at time of focal post (general FB behavior)
Mean_time_photos	Average time between photo uploads (general FB behavior)
Mean_time_videos	Average time between video uploads (general FB behavior)
Mean_time_links	Average time between links (general FB behavior)
Mean_time_likes	Average time between user likes (general FB behavior)
Mean_time_posts	Average time between user posts (general FB behavior)
SD_time_photos	Stand. deviation of the time between photo uploads (general FB behavior)
SD_time_videos	Stand. deviation of the time between video uploads (general FB behavior)
SD_time_links	Standard deviation of the time between links (general FB behavior)
SD_time_likes	Standard deviation of the time between user likes (general FB behavior)
SD_time_posts	Standard deviation of the time between user posts (general FB behavior)
Profile_completeness	Number of Facebook profile items filled in by the user (general FB behavior)

Table A3
Lagging variables.

Variable name	Variable description
Nbr_likes	Number of likes the focal post received in 7 days
Nbr_comments	Number of comments the focal post received in 7 days
Nbr_own_comm	Number of comments made on the focal post by the focal user
Nbr_comm_persons	Number of persons commenting on the focal post
Nbr_comm_likes	Number of likes on comments received on the focal post
Nbr_words_comm	Number of words in the comments received on the focal post
Nbr_punct_comm	Number of punctuations in comments received on the focal post
Nbr_qm_comm	Number of question marks in comments received on the focal post
Nbr_em_comm	Number of exclamation marks in comments received on the focal post
Nbr_upper_comm	Number of uppercase letters in comments received on the focal post
Nbr_lol_comm	Number of 'lol' mentioned in comments received on the focal post
Neg_emo_comm	Number of negative emoticons in comments received on the focal post
Pos_emo_comm	Number of positive emoticons in comments received on the focal post
Dev_nbr_likes	Deviation in the number of likes received on the focal post compared to previous posts
Dev_nbr_comments	Deviation in the number of comments received on the focal post compared to previous posts
Dev_nbr_own_comm	Deviation in the number of own comments made on the focal post compared to previous posts
Dev_nbr_comm_persons	Deviation in the number of commenting persons on the focal post compared to previous posts
Dev_nbr_comm_likes	Deviation in the number of likes received on comments on the focal post compared to previous posts
Dev_neg_emo	Deviation in the number of negative emoticons in comments received on the focal post compared to previous posts
Dev_pos_emo	Deviation in the number of positive emoticons in comments received on the focal post compared to previous posts
Comments_span	The time span in which comments were received

Appendix B: Variable importance scores

Table B1
Variable importances (top 50).

Rank	5*2 CV median mean decrease in Gini	Variable name	Category
1	159	Number_uppercase	Focal post's variables
2	150	Nbr_likes	Lagging variables
3	135	Neg_emo_comm	Lagging variables
4	134	Dev_neg_emo	Lagging variables
5	117	Dev_nbr_likes	Lagging variables
6	109	Month	Focal post's variables
7	69	Deviation_neg_words	Leading variables
8	69	Polarity	Focal post's variables
9	67	Mean_neg_emo	Leading variables
10	61	Deviation_polarity	Leading variables
11	56	Mean_pos_emo	Leading variables
12	45	Number_punctuation	Focal post's variables
13	45	Age	Leading variables

14	43	Number_neg_words	Focal post's variables
15	42	Dev_nbr_comments	Lagging variables
16	42	Dev_pos_emo	Lagging variables
17	38	SVD Concept 1	Focal post's variables
18	37	SVD Concept 22	Focal post's variables
19	35	Weekday	Focal post's variables
20	35	SVD Concept 29	Focal post's variables
21	34	SVD Concept 2	Focal post's variables
22	33	Mean_likes_posts	Leading variables
23	32	Nbr_comm_persons	Lagging variables
24	32	SVD Concept 62	Focal post's variables
25	31	Nbr_words_comm	Lagging variables
26	31	Total_nbr_likes	Leading variables
27	31	SVD Concept 21	Focal post's variables
28	31	SVD Concept 28	Focal post's variables
29	30	SVD Concept 99	Focal post's variables
30	30	Mean_time_likes	Leading variables
31	29	SVD Concept 48	Focal post's variables
32	29	Deviation_subjectivity	Leading variables
33	29	SVD Concept 10	Focal post's variables
34	29	Mean Polarity	Leading variables
35	29	SVD Concept 6	Focal post's variables
36	29	SVD Concept 81	Focal post's variables
37	29	SVD Concept 34	Focal post's variables
38	28	SVD Concept 78	Focal post's variables
39	28	Number_characters	Focal post's variables
40	28	SVD Concept 13	Focal post's variables
41	28	SVD Concept 83	Focal post's variables
42	28	SVD Concept 9	Focal post's variables
43	28	SVD Concept 3	Focal post's variables
44	28	SVD Concept 25	Focal post's variables
45	28	SVD Concept 63	Focal post's variables
46	28	SVD Concept 53	Focal post's variables
47	28	SD_time_likes	Leading variables
48	28	SVD Concept 18	Focal post's variables
49	28	SVD Concept 7	Focal post's variables
50	28	Nbr_comments	Lagging variables

3

Linking Event Outcomes to Customer Lifetime Value: The Role of MGC and Customer Sentiment

3. Linking Event Outcomes to Customer Lifetime Value: The Role of MGC and Customer Sentiment

Abstract

Concurrent with firms' expanding investments in social media, aimed at monitoring or driving engagement, marketing executives continue to express concerns regarding the effectiveness of these investments in impacting performance outcomes. Specifically, linking customers' experiences to individual-level performance metrics, and the attribution of marketing levers, remain key challenges. In this study, we explore the moderating role of Marketer Generated Content (MGC) on the customer experience (measured objectively by event outcomes) -- customer sentiment relationship, and further demonstrate the importance of customer sentiment for modeling customers' direct engagement with a firm. We demonstrate the ability of MGC to attenuate the negative effect of negative customer experiences on direct customer engagement measured based on an individual's CLV. Based on a series of counterfactual analyses, we further demonstrate the relative tradeoff in customer sentiment improvements based on increasing the volume of MGC surrounding particular experience encounters versus achieving perfect performance. The results of these analyses suggest that managers may find greater potential returns from increasing their investments in MGC as opposed to focusing exclusively on improvements in performance during individual service encounters, and that additional MGC posting efforts may be particularly effective in lifting consumer sentiment based on encounters with neutral or negative performance.

This chapter is based on Meire, M., Hewett, K., Ballings, M., Kumar, V. and D. Van den Poel (2018), working paper, Ghent University. To be submitted to Journal of Marketing.

1. Introduction

In response to the explosion in potential customer-brand touch points and the increasingly social nature of customers' experiences (Lemon and Verhoef, 2016), firms' investments in social media (SM) continue to expand; U.S. firms' investments are projected to reach nearly $17.5 billion by 2019 (Statista, 2018). Among the top firm uses of SM are social listening, or monitoring customers' sentiment related to their experiences (Caruso-Cabrera and Golden, 2016) and contributing content aimed at managing perceptions, such as for customer care or service recovery purposes (Ma et al., 2015), or for driving engagement (Goh et al., 2013; Harmeling et al., 2017). Despite this upward trend in SM marketing and monitoring, however, executives continue to express doubt regarding the effectiveness of these investments (Moorman, 2017; Stein, 2016). In addition, the attribution of marketing levers (Marketing Science Institute, 2016) and linking customers' experiences to individual-level performance metrics (KPMG, 2016) remain key challenges for managers.

Whereas academic research has demonstrated the importance of customer sentiment in SM for assessing customer perceptions of experiences (Micu et al., 2017) or products (Babić Rosario et al., 2016), studies incorporating objective data regarding customers' actual brand-related experiences is rare. This is surprising given the ability of firms in many contexts to monitor objective characteristics of their offerings in real-time. For example, customer contact centers track metrics such as response or resolution time, and wait time in a queue (Rongala, 2016); and retailers can monitor performance variables such as offline store traffic patterns, wait-times, inventory, and checkout times (Stores.org, 2017). We view marketers' abilities to influence sentiment and drive customer engagement (CE) as a result of actual customer experiences as an un-tapped use of SM and an under-researched area in marketing.

This study aims to shed light on the role of firms' SM investments (called marketer generated content (MGC) hereafter) in managing customer perceptions of their actual experience encounters, and ultimately driving their direct engagement with a firm. Our research addresses three primary questions:

1) Can marketers influence customer perceptions of actual experiences via their SM contributions? That is, can marketers temper negative sentiment and magnify positive sentiment regarding actual experience above and beyond the characteristics of the experiences themselves? If so, what are the optimal conditions for these posts in terms of volume?

2) Does customer sentiment in SM drive direct CE? If so, can marketers enhance direct engagement via their own SM contributions?

3) Aside from customer sentiment, what is the role of other SM variables such as page likes and network characteristics in driving direct CE?

To answer these questions, we built an unprecedented longitudinal database featuring brand-related customer-level SM activity metrics including liking a brand's SM page, MGC, likes of and comments on brand posts, RSVPs for events sponsored by the brand, and features of customers' networks all linked to transaction variables at the customer level. Importantly, we also capture other brand communication variables as well as various objective characteristics of a particular customer experience encounter, enabling us to assess the moderating impact of MGC on the customer experience—customer sentiment relationship. The chosen context for this study is the Facebook fan page of a European soccer team. This context enables us to capture variance in experiences across interactions due to differences in attributes such as opposing teams and their attending fans, with regular opportunities for interaction with the brand via matches, and creates ample opportunity to observe experience-related customer sentiment based on fans' frequent SM use to discuss sports (Catalyst, 2013). Whereas empirical evidence from the marketing literature is mixed regarding the effectiveness of MGC for behavioral outcomes such as purchases (Xie and Lee, 2015), considered a form of direct CE, we conclude that marketers can provide cues to influence reactions to actual customer experiences without influencing objective characteristics of the experiences themselves, and in doing so can positively influence the impact of those experiences on customer sentiment in SM. Moreover, we show that customer sentiment is positively related to direct CE, with a larger relative effect on purchase probability compared to contribution margin, and that MGC can thus indirectly influence direct CE through customer sentiment. Finally, our results show that SM measures such as page likes and the number of SM interests, are not related to direct CE.

2. Review of Relevant Literature

As firms have begun to see the importance of SM for customer relationship management, research has begun to expand our understanding of the impact of SM marketer generated content (MGC) on important outcomes such as individual customer behavior, and product, brand, or firm-level performance. In addition, an increasingly rich body of work has focused on customer sentiment in SM, examining the influence of user-generated content (UGC) at an aggregate level on individual customer behaviors or product, brand, or firm-level performance.

For example, aggregate-level UGC such as product reviews or tweets mentioning a particular product or brand may impact individual behavior, product sales, or firm performance. Research focusing on individual-level contributions such as a customer's posts on or engagement with a brand's Facebook page has also demonstrated links to important customer behaviors and have accounted for a variety of important individual characteristics. Whereas most studies in this domain focus on particular forms of customer or firm sentiment, such as the valence or volume of SM posts, or other forms of SM contributions such as likes or sharing posts, accounting for multiple forms of either UGC or MGC would enable researchers to capture more completely the complexities of the SM environment in which both customers and firms operate. Surprisingly, studies incorporating multiple forms of either UGC or MGC in a single research setting are rare.

Aside from the contributions of SM participants, another factor influencing reactions and contributions to the SM environment is customers' actual experiences with marketers' goods and services. While research in this domain offers evidence regarding the importance of product features such as price or quality (Chen et al., 2011; Nam et al., 2010), or buyer-seller relationship characteristics such as tenure (Kumar et al., 2016) for customer behaviors, studies investigating these relationships in association with a particular brand or firm interaction are scarce. That is, studies tend to examine UGC and/or MGC over a particular period of time without aligning to a particular encounter. This is surprising given the fact that many marketers are able to track individual customers' offline interactions such as their presence at events and appointments, or completion of a service, as well as evidence in the literature regarding the importance of customers' experiences with a firm or brand for loyalty behaviors (Lemon and Verhoef, 2016). Indeed, marketers commonly design marketing communications to coincide with those interactions, such as sending reminders or posting promotional content online. In this study, we assess the moderating role of MGC on the link between objective characteristics of identifiable customer experience encounters and customer sentiment in SM, and further link customer sentiment to direct CE, captured via their Customer Lifetime Value (CLV) (Pansari and Kumar, 2017). In addition, we account for multiple forms of both UGC and MGC.

We highlight prior relevant studies in comparison with the present study in Table 3.1. In particular, we focus in Table 3.1 on studies examining UGC or MGC in association with individual-level behavioral outcomes that enable modeling of direct customer engagement as measured by CLV. On the basis of the comparisons offered in Table 3.1, we summarize here two primary contributions of our study: 1) We demonstrate that marketers, via their MGC on

SM, can temper negative customer sentiment and magnify positive sentiment regarding actual experiences, which we then link to direct customer engagement. Based on the connection between customers' direct CE and firm value (Kumar, 2018), this tie to engagement is an extremely important issue for marketers in their attempts to demonstrate value from their SM investments. In addition, by accounting for multiple forms of UGC (page likes and comments) and MGC (SM posts and emails) in assessing these relationships, we attempt to capture more fully the complexities in the SM environments in which firms and customers interact, and to further extend research in this domain. In establishing these relationships we also respond to the call from numerous scholars to include WOM in models assessing customer value (Hogan et al., 2003; Kumar et al., 2010a; Libai et al., 2010). 2) We examine the influence of objective performance characteristics of an individual customer experience encounter on customer sentiment and assess the moderating influence of MGC on this relationship. In doing so we are able to offer direction to marketers in crafting targeted communications that have the potential to enhance direct CE based on characteristics of actual customer experience encounters. In addition, we offer insight into how customer touch points influence behavioral outcomes such as loyalty (Lemon and Verhoef, 2016), and address calls for further research on MGC in SM (Kumar et al., 2016).

3. Conceptual Framework

In Figure 3.1, we provide our conceptual framework. As depicted in the figure, we aim to investigate the relationships among customer experiences, customer sentiment, marketer SM content, and direct CE. We leverage customer engagement theory (Pansari and Kumar, 2017) in establishing the relationships in our conceptual framework, discussed below. We begin by conceptualizing our key dependent variables, and then describe the expected relationships in our framework.

3.1. Customer Engagement

The concept of CE as defined by Pansari and Kumar (2017), and consistent with (Kumar et al., 2010a), is defined as the process by which a customer adds value to the firm, either through direct or/and indirect contribution, with direct contributions consisting of customer purchases, and indirect contributions referring to the customer's incentivized referrals, brand-related SM conversations, and feedback to the firm. Whereas direct contributions can be linked to important customer-level value metrics such as CLV (Pansari and Kumar 2017), indirect contributions include behaviors such as SM WOM. In this study, we focus on direct CE as our key outcome

Citation	Research Focus	Included/accounted for in model:				
		Multiple forms of MGC	UGC			Actual brand experience (objective perf.)
			Multiple forms	by focal customer	by others	
THIS STUDY	**importance of MGC in shaping customer sentiment surrounding a particular experience encounter, and the value of customer sentiment for customer engagement**	**X**	**X**	**X**	**X**	**X**
John et al.(2017)	whether "liking" a brand influences brand evaluations	X		X	X	
Mochon et al. (2017)	how Facebook page likes affect offline customer behavior			X		
Baker, Donthu, and Kumar (2016)	how WOM conversations about a brand relate to purchase & retransmission intentions			X		
Kumar et al. (2016)	MGC's effect on customer behavior & profitability	X		X		
Beukeboom, Kerkhof, and de Vries (2015)	whether following a brand's FB updates can cause positive changes in brand evaluations			X		
Homburg, Ehm, and Artz (2015)	consumer reactions to firm participation in C-to-C online conversations			X	X	
Manchanda, Packard, and Pattabhiramaiah (2015)	effect of customers joining a firm's SM community on expenditures		X	X		
Xie and Lee (2015)	effects of exposures to earned & owned SM activities on brand purchase	X			X	
Kumar et al. (2013)	How SM can be used to generate positive WOM & influence performance	X	X	X	X	
Goh, Heng, and Lin (2013)	impacts of both UGC & MGC on repeat purchase behaviors	X	X	X	X	
Rishika et al. (2013)	effect of customers' participation in firm initiated SM efforts on customer value		X	X	X	
Nam, Manchanda, and Chintagunta (2010)	effect of service quality on customer acquisition, accounting for spillover effects from WOM				X*	X

* Not modeled directly, rather, assumed based on geographic proximity to other subscriber

Table 3.1: Comparison with relevant literature

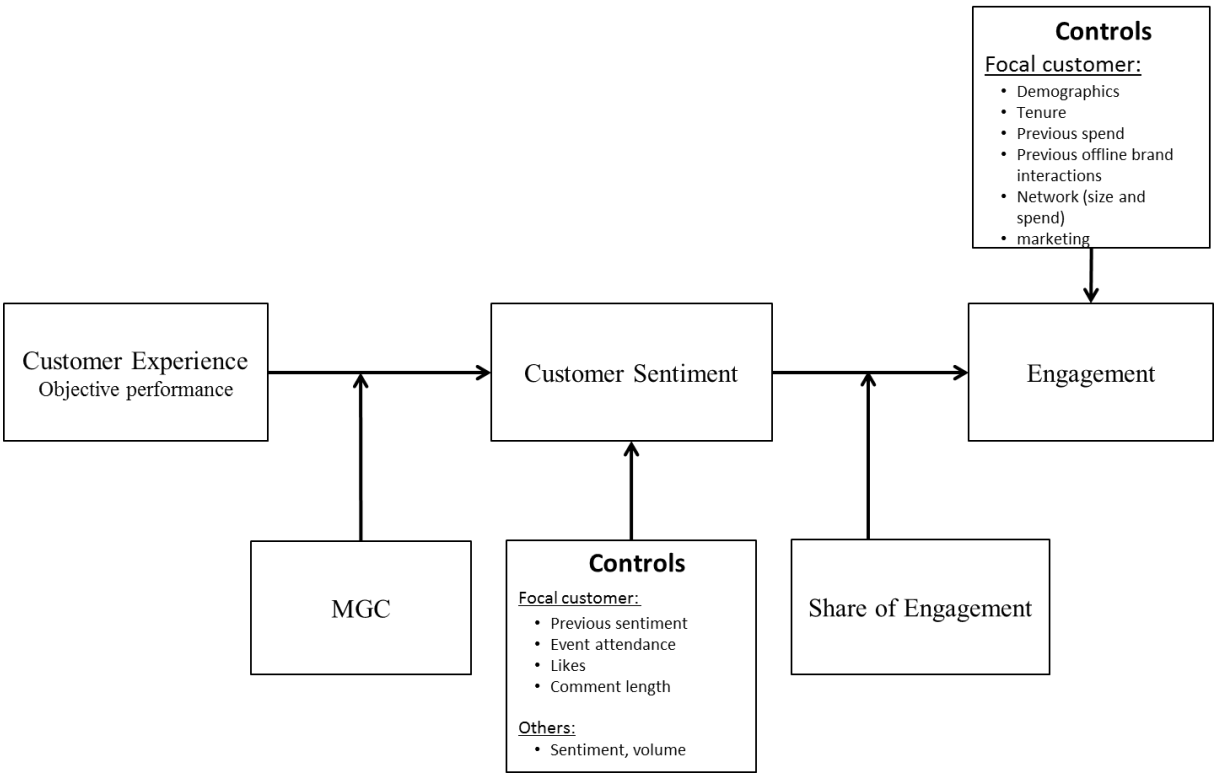


Figure 3.1: Conceptual framework

variable due to its impact on firm value (Gupta et al., 2004), its importance for developing a competitive advantage (Brodie et al., 2013), driving sales growth (Voyles, 2007), and helping firms allocate resources efficiently (Kumar et al. 2008).

The academic literature has begun to recognize firms' efforts to actively influence CE, and to link these efforts to customers' experiences. Harmeling et al. (2017) argue that marketers' actions can enhance the effect of customers' experience on CE based on its ability to strengthen existing cognitive bonds by encouraging task-based activities such as posting comments. Such activities, as part of a firm's CE marketing efforts, are aimed at influencing customers beyond their experiences with the firm's offerings. Whereas Harmeling et al. (2017) specifically identify CE initiatives as activities that require active participation, such as campaigns that urge customers to complete a task or participate in an event, we propose another route through which marketers can influence CE, namely, by enhancing the likelihood of positive customer sentiment surrounding individual brand-related experiences.

3.2. Understanding Customers' Experiences

Customer experience (CX) has been conceptualized as a multidimensional construct with cognitive, emotional, behavioral, sensorial, and social components (Schmitt, 1999, 2003; Verhoef et al., 2009) based on customer-firm contacts that occur at distinct points in time, called

touch points (Homburg et al., 2015; Schmitt, 2003). With regard to services specifically, CX is typically conceptualized as the customer's direct and indirect experience of the service process, the organization, and the facilities, as well as how the customer interacts with the firm's representatives and other customers, all of which in turn generate customers' cognitive, emotional, and behavioral responses (Chahal et al., 2015). We measure customer experiences based on the objective event outcomes.

There is evidence that customers' SM contributions can be valuable in providing insight into CX. For example, research has shown the ability of UGC on SM to capture brand perceptions (Schweidel and Moe, 2014), identify service intervention opportunities (Ma et al., 2015), and predict buyer behavior (Baker et al., 2016; Trusov et al., 2009), and to assess customer perceptions of experiences (Micu et al., 2017) or products (Babić Rosario et al., 2016). SM content can also deliver more timely feedback regarding CX (Luo et al., 2012) relative to other tools such as surveys that typically involve response delays, and is real-time relevant in terms of content (Lemon, 2016). In the information systems literature, related studies focus on demonstrating the use of methods such as quantitative analysis, text mining, and sentiment analysis to analyze UGC in SM (Farhadloo et al., 2016; He et al., 2016; Misopoulos et al., 2014), asserting that such feedback reflects CX. In this study we empirically expose the value of the sentiment in customer SM comments for CE, and further demonstrate the role of MGC, a form of CE marketing (Harmeling et al. 2017), to moderate the influence of actual experiences on customer sentiment.

3.3. Customer Experiences and Sentiment

The academic literature offers ample evidence that customer sentiment in SM can describe customers' experiences. For example, research has demonstrated the value of customers' online reviews for identifying the CE factors that influence overall satisfaction (Farhadloo et al., 2016; He et al., 2016; Misopoulos et al., 2014), which has been linked to individual-level outcomes such as spending growth (Fornell et al., 2010), and willingness to pay (Homburg et al., 2005). Because customer-initiated SM content is argued to be more customer centric relative to approaches in which marketers determine what is asked in gathering feedback (Villarroel Ordenes et al., 2014), customers' SM comments may be inherently more reflective of the true nature of their experiences. In addition, such comments may be naturally more emotional since the customer was motivated to provide them without the firm's prompting. According to Pansari and Kumar (2017), customer-firm relationship quality depends in part on the customer's level of emotional connectedness toward the relationship. Consistent with this perspective, we expect

customer-initiated SM comments to reflect their reactions to their brand-related experiences. Research demonstrating the use of sentiment analysis of customers' SM contributions to uncover service experiences (Misopoulos et al., 2014) also supports this expectation.

3.4. The Moderating Role of MGC

As noted above, marketers' contributions to SM can encourage active CE. We further argue that these contributions can influence customer sentiment based on actual experiences. There is evidence in the marketing literature to support this expectation. First, van Doorn et al. (2010) argue that the value of MGC for CE may depend on contextual factors such as customers' satisfaction with firm interactions. In addition, there is evidence that customers' reactions to their experiences may be more malleable than the satisfaction literature has typically assumed. For example, Pham et al. (2010) find that contextual cues that increase customers' self-awareness can influence their reactions to experiences with service providers. Research has also documented the ability of more traditional marketing actions such as feature advertising and promotional in-store displays (Ngobo, 2017) and digital advertising message content (Bruce et al., 2017) to influence customer transition across loyalty conditions. Thus, without changing the objective performance of the service itself, there is evidence in support of marketers' abilities to influence customer reactions.

Recognizing this ability to influence reactions to actual brand experiences, practitioners have begun directing some of their SM investments to identify moments or points at which the firm can influence the customer's overall "journey" with the brand. For example, Starwood Hotels texts guests with information based on their interactions, such as turning their cell phones into virtual keys as they approach their rooms, enabling them to open them via an app, and also sends well-timed dining recommendations (Edelman and Singer, 2015). The rise of the Chief Customer Experience Officer position, typically charged with breaking down silos to blend marketing and CX with the goal of maximizing direct CE (Kokes, 2017) also support this expectation. Thus, through various methods, firms are increasingly focused on developing practices to connect with customers at more of an emotional level to make meaningful interactions and enhance the potential impact of their brand experiences (Taparia, 2015). From an engagement theory perspective, this enhanced emotional connection with customers should lead to greater direct engagement.

3.5. Customer Sentiment and Direct Customer Engagement

The link between customers' perceptions of their experiences and behavioral outcomes such as purchases is well established in the academic literature. Studies have confirmed the positive influence of satisfaction, an outcome of positive CX perceptions, on purchase behavior (Bolton, 1998), consistent with the notion of the service-profit chain (Anderson and Mittal, 2000). In addition, satisfaction is also argued by Pansari and Kumar (2017) to be one of the key requirements for CE. Thus, we anticipate customer sentiment in SM, as it is reflective of their experiences and the event outcomes, to be positively associated with their direct engagement (CLV) with the firm. Furthermore, we expect this relationship to hold above and beyond the influence of other relationship factors such as customers' previous relationship activities with the firm, which have been demonstrated in previous studies to be valuable in estimating customers' direct engagement with firms.

3.6. Moderating Impact of Share of Interests

We expect the impact of customer sentiment on CLV to be moderated by the share of interests the firm receives in the SM environment. Previous work has found similar notions such as share of wallet to influence CLV (Reinartz, Thomas, and Kumar 2005), which we translate to the SM domain. Consistent with arguments that firms that own a greater share of their customers' wallets enjoy stronger relationships in general based on characteristics such as greater relationship duration and an enhanced ability to learn about customer needs via more communication (Anderson and Narus 2003), we expect a positive moderating influence of share of interests. The greater the marketer's share of customers' interests (the smaller the overall number of interests) in SM, the greater will be their attention in SM in general on issues related to the firm or brand, and the more meaningful their firm- or brand-related comments in SM are likely to be in terms of their behaviors, i.e., their CLV.

Next, we describe our context and data, and then detail the modeling approach we followed in answering our research questions. Then we discuss our results and important theoretical and managerial implications of our work. We conclude with a discussion of limitations and future research directions.

4. Data

We focus in this study on the dominant SM platform: Facebook (John et al., 2017). More specifically, the context for this study is the Facebook fan page of a European soccer team. In order to extract the data from Facebook, we employed two options. First, in order to model customer sentiment, we extract all comments on team posts on its official Facebook page,

resulting in a total of 265,530 data points (comments) from 52,431 Facebook users. This includes all comments from the period 2011-2015, which represent all the seasons after the team's Facebook fan page was established in 2010. In addition, we also include likes on team posts and declared attendance at team events, as indicated on Facebook, information publicly available using the Facebook API. Since we evaluate customer sentiment at the user—comment level per match, we also include data on match conditions and outcomes (the objective performance characteristics) as well as the team's SM contributions (Facebook posts) during a +two day-window of a match (described in further detail below).

Second, in order to link customer sentiment to direct CE, we developed an application that enables us to collect buyer information for matching with the transactional database. The app was hosted on the team's Facebook page as well as the team's main page tabs, and advertised on the team's main page four times over a period of three weeks. To encourage usage of the app we offered a chance to win a prize (shirt signed by a player) to participants. Once users clicked on a link provided in the ad, they were presented with an authorization box in which they had to give their permission before any data were gathered, and were told exactly what would be gathered. Once opened, the app presented an activity including three team-related questions and one tie breaker (how many contest participants) to determine a winner. Meanwhile, the app collected user demographics (age, gender, location) and SM information (page likes, comments and likes on posts, user posts and declared event attendance). Data were collected between May 7 and May 26, 2014 and go back until August 9, 2007. We gathered data on 1,107,222 Facebook users. We provide further details regarding the data in our discussion of the modeling and additional analyses, below.

We merged customers' Facebook data with the soccer club's customer database via either name, city, and age or e-mail address. The database includes transactional information (e.g., frequency of tickets sold, monetary value), customer specific information (e.g., name, gender, birthday, city, email), relationship information (e.g., openness to team emails, number of team emails sent, email click rate), and behavioral information (matches attended). A total of 28,131 out of 89,797 customers were matched via this procedure; however, we limit our investigation to those who purchased at least one season ticket during the 48-month period 2011-2015. Season tickets represent the vast majority (on average 83%) of the team's total ticket sales. We focus on the period 2011-2015 since the team's official Facebook page was established during the 2010 season. We identified a total of 24,341 customers who bought at least one season ticket in this period. Finally, we selected customers that have used the application and made at least

one comment on the Facebook page during the 48-month window; we ended up with 5,783 matched customers.

5. Model Descriptions

5.1. Customer Sentiment Model

We model customer sentiment for Facebook users who commented on the team's Facebook posts[1]. We model customer sentiment as each user's online expressed sentiment across all matches, over 48 months, yielding a total of 212 potential experience encounters per customer. Our rationale for using these comments is as follows. First, there is apparent agreement that information contained in SM can be useful for monitoring customer sentiment (Luo et al., 2012), and that it can capture the dynamic nature of customers' experiences more effectively than periodic survey approaches (Chien et al., 2016; Moe and Trusov, 2011). Text-based SM content is also argued to be more valuable than numerical ratings or volume metrics, which ignore information contained in comments (Tirunillai and Tellis, 2012). Because their timing and content are driven by the users themselves, SM comments may also be less susceptible than surveys to issues such as memory effects. While some researchers have argued that UGC can differ based on the platform (Schweidel and Moe, 2014), Smith et al. (2012) find that positive brand-related sentiment in UGC does not differ across user platforms (Facebook, Twitter, and Youtube) and argue that brand experiences in particular influence comments on Facebook, which makes it the most suitable platform for our research.

We restrict customer sentiment measurement to comments on the team's Facebook posts within a +two day-window starting from the end of a match, in order to: 1) increase the likelihood that comments are related to a particular interaction experience, thereby reducing "noise" in the way of comments unrelated to team interactions; 2) reduce the likelihood of capturing comments regarding multiple matches, since they are at times as close as four days apart. A similar approach could be used in other contexts, with windows around specific interactions, such as purchases instances. We restrict the explanatory variables to information that is available before the focal comment (dependent variable) is made (e.g., we only include the number of team posts posted before the focal comment was made), which, in combination with our approach to only include comments after the match, alleviate endogeneity concerns.

[1] There are no possibilities for users to post messages on this Facebook page other than the possibility to react to the team's posts.

We use a classification algorithm based on a sentiment lexicon to determine sentiment for comments on posts (Goh et al., 2013). Next, based on evidence that the effect of neutral comments is much smaller in magnitude than that of both positive and negative comments (Sonnier et al., 2011), and arguments that positive and negative comments are most relevant for extracting opinions, review polarity (Godes and Mayzlin, 2004), or sentiment (Tirunillai and Tellis, 2012), we proceed with only positive and negative comments. This results in a total of 44,206 user-match records (for 18,075 users) used for our model. We use a generalized linear mixed effects model (with a logistic link function) to model customer sentiment. Next, we discuss our independent variables.

First, we look into match-specific variables, which reflect the objective performance measures. First, we include the match result (win, loss, or draw; $Result_m$), since this is the principal appeal of watching sports (Madrigal, 1995) and could be expected to positively influence customer sentiment (Van Leeuwen et al., 2002). In addition, match quality and outcome are the two most important attributes in evaluating service quality in a sports context (Kelley and Turley, 2001), an aspect of CX. We include three variables to reflect match quality: opponent quality, match type (defined below) and number of red and yellow cards. Playing against better teams in more intense circumstances (e.g., a cup final) implies a higher quality match (Cyrenne, 2001), and leads to greater BIRG and enjoyment (Madrigal, 1995). Opponent quality is based on the previous year's results and team-specific arguments (e.g., a derby) as a categorical variable with three levels (1=low; 2= medium; 3=high; $QualityOpponent_m$). Match type is a categorical variable with four levels (1=cup, 2=competition, 3=competition play-offs or 4=European cup; $TypeMatch_m$). Finally, we include the number of yellow and red cards, given for moderate or very severe fouls by the focal team's players, respectively ($YellowCards_m$ and $RedCards_m$). Cards contribute negative outcomes (Castellano et al., 2012), a negative touchpoint (Funk, 2017) and lower match quality. We expect a negative relationship between number of cards and customer sentiment.

Next, in order to investigate the moderator effect of MGC on the relationship between the objective CX and customer sentiment, we include the number of posts on the Facebook page of the team ($MGC_{u,c,m}$), and the interaction effect between the result of the match and MGC. We expect MGC to positively influence customer sentiment.

We leverage the CX and sports satisfaction literatures to identify relevant control variables for customer sentiment. First, team identification is the extent to which individuals perceive themselves as fans of and involved with a team, care about its performance and view it as a

representation of themselves (Branscombe and Wann, 1992)). A related concept is Basking In Reflected Glory (BIRG), or sharing in the glory of a successful other (Cialdini et al., 1976). Both concepts are related to social identity and self-categorization theories (Tajfel and Turner, 1979; Turner et al., 1987), which suggest that group identification occurs when a social category is relevant and important to individuals, and group actions are central to their social identities (Wann, 2006). Team identification and BIRG can amplify repatronage intentions and match experience (Clemes et al., 2011; Wakefield, 1995).

Participation in team-related SM can be seen as a form of online team identification. If a user likes a post or page, this appears in the news feed of his/her friends who then associate him/her with the "liked" company. Therefore, we include the number of likes for team Facebook posts during the match window ($likesMGCPosts_{u,m}$) and intentions (yes/no) to attend the event ($EventFacebook_{u,m}$). More likes or declarations to attend a match increase team identification, which we expect to be positively related to customer sentiment.

Next, we include a dummy variable to indicate actual attendance ($EventAttending_{u,m}$), and expect it to positively influence experience, as the customer can feel the atmosphere to a greater extent than watching it on television or online. We include previously expressed customer sentiment ($Customer\ Sentiment_{u,c,m-1}$) based on comments during the previous match window for which the user commented, as this is a frequently mentioned as a predictor of CX (Lemon and Verhoef, 2016; Verhoef et al., 2009). Thus, for first-time comments, no previous sentiment is available, and the value is zero. We include previous comment volume in the post's thread ($OtherUGCVolume_{u,c,m}$), and expect a negative relationship with the focal comment's sentiment (Moe and Trusov, 2011). We approximate the context of the comments by including the valence of the last comment in the thread before the focal comment (hence we lose the first comment per thread; $OtherUGCValence_{u,c,m}$), and we expect a positive relationship with customer sentiment (Homburg et al., 2015; Moe and Trusov, 2011). Finally, we include comment length, measured as the logarithm of the word count ($Comment\ Length_{u,c,m}$), based on Homburg et al. (2015), and expect a negative relationship with customer sentiment. The customer sentiment equation takes the following form:

$$Customer\ Sentiment_{u,c,m} = \alpha_0 + \alpha_{1,u} + \alpha_{2,m} + \alpha_3\, Result_m + \alpha_4\, RedCards_m + \quad (1)$$
$$\alpha_5\, YellowCards_m + \alpha_6\, TypeMatch_m + \alpha_7\, QualityOpponent_m +$$
$$\alpha_8\, MGC_{u,c,m} + \alpha_9\, MGC_{u,c,m} * Result_m + \alpha_{10}\, \theta_m + \alpha_{11}\ Home\ Game_{u,m} +$$
$$\alpha_{12}\, likesMGCPosts_{u,c,m} + \alpha_{13}\ EventFacebook_{u,m} +$$
$$\alpha_{14}\, EventAttending_{u,m} +$$
$$\alpha_{15}\, Customer\ Sentiment_{u,c,m-1} + \alpha_{16}\, OtherUGCValence_{u,c,m} +$$
$$\alpha_{17}\, OtherUGCVolume_{u,c,m} + \alpha_{18}\, Comment\ Length_{u,c,m} + \varepsilon_{u,c,m},$$

where $Customer\ Sentiment_{u,c,m}$ denotes the customer sentiment for user u, as expressed in comment c during match m and $\varepsilon_{u,c,m}$ is the error term. The variable θ_m represents a vector of year dummies accounting for factors that may vary by year. The variables *Result*, *TypeMatch* and *QualityOpponent* are also operationalized as a vector of dummies.

Model Free Evidence

In Figure 3.2 we plot the average customer sentiment related to wins, losses and draws (objective performance), for different levels of MGC (low, medium and high). The figure provides model-free evidence of the proposed relationship between customer sentiment and the outcome of the experience, and preliminary evidence that MGC is able to moderate this relationship, especially in the case of draws and losses.

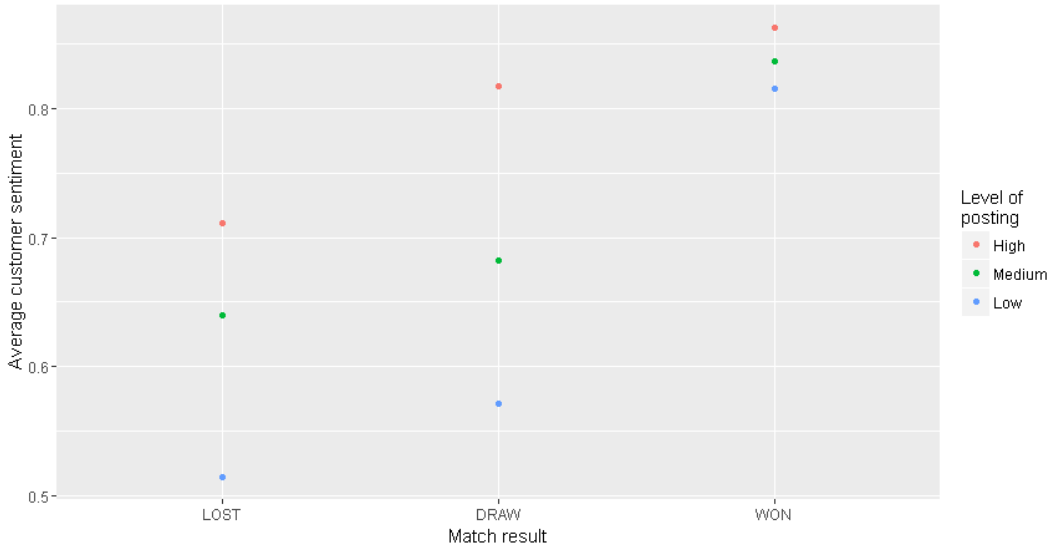


Figure 3.2: Model-free evidence of the relationship between objective performance, customer sentiment and MGC

Table 3.2: Overview of the variables for customer sentiment modeling

Variable	Description	Variable origin
Dependent	*Measure of subjective sentiment wrt performance*	
CustomerSentiment $_{u,c,m}$	Dependent variable. Customer *u* sentiment, as expressed in comment *c*, during match *m*. (binary variable)	(Tirunillai and Tellis, 2012)
Customer Experience	*Objective measures of service performance*	
Result (Lost)$_m$	Dummy variable indicating whether the match *m* was lost by the focal team (in contrast to a draw)	Madrigal (1995), Van Leeuwen, Quick and Daniel (2002)
Result (Won)$_m$	Dummy variable indicating whether the match *m* was won by the focal team (in contrast to a draw)	
RedCards$_m$	The number of red cards for the focal team in match *m*	Funk (2017), Castellano, Casamichana and Lago 2012
YellowCards$_m$	The number of yellow cards for the focal team in match *m*	
TypeMatch (Eur)$_m$	Dummy indicating whether match *m* is a European match	Kelley and Turley (2001), Cyrenne (2001)
TypeMatch (Nor)$_m$	Dummy indicating whether match *m* is a normal match in competition	
TypeMatch (PO)$_m$	Dummy indicating whether match *m* is Play-Off match at the end of the competition (which may be of higher intensity)	
QualityOpponent (Medium)$_m$	The opponent in match *m* is of a medium level	Kelley and Turley (2001), Cyrenne (2001)
QualityOpponent (High)$_m$	The opponent in match *m* is of a high level	
MGC	*Measures of Marketer Generated Content*	
MGC $_{u,c,m}$	Number of posts on Facebook by the focal team between the end of match *m* and the time of posting of comment *c* by user *u* (MGC volume)	Homburg, Ehm & Artz (2015), Goh, Heng and Lin (2013)
Result $_m$ * *MGC* $_{u,c,m}$	Interaction effect between Result of match *m* and MGC	
Control variables	*Control variables for customer sentiment*	
ϑ_m	Dummy variables indicating the year in which the match *m* was held	
Home Match$_m$	Dummy indicating whether the match *m* is a home match	
Likes MGC Posts$_{u,c,m}$	The number of likes of user *u* on posts of the focal team during the timeframe of match *m* (logarithm)	Wann (2006), Clemes, Brusch and Collins (2011)
EventFacebook$_{u,m}$	Dummy indicating whether user *u* has declared on Facebook to attend match *m*	Wann (2006), Clemes, Brusch and Collins (2011)
EventAttending $_{u,m}$	Dummy indicating whether user *u* has attended the match *m* (from soccer team database)	
Customer Sentiment $_{u,c,m-1}$	Lag of measured customer sentiment of user *u*	Lemon and Verhoef (2016)
Other UGC Valence $_{u,c,m}$	Valence of the previous comment in the post thread of comment *c* by user *u* during match *m*	Moe and Trusov (2011), Homburg, Ehm and Artz (2015)
Other UGC Volume $_{u,c,m}$	Volume of other user's comments in the post thread of comment *c* by user *u* during match *m*	Moe and Trusov (2011)
Comment length $_{u,c,m}$	Length of the comment *c* by user *m* during match *m* (logarithm)	Homburg, Ehm and Artz (2015)

See Table 3.2 for a list of variables in the customer sentiment equation and Appendix A for measures of variables in this equation, the distribution of our dependent variable and the correlation matrix of non-categorical variables, indicating no multicollinearity issues.

5.2. Engagement model Specification

Self-selection issue. It is possible that our data suffer from sample selection bias as customers included in the engagement-analysis self-selected into this study by allowing us to extract their Facebook information via the app. These individuals may not be representative of the population as there may be unobserved factors that influence both the decision to use the application and buying behavior (and hence engagement). This self-selection potentially leads to an endogeneity issue due to omitted variables bias (Wies and Moorman, 2015) which is alleviated by implementing a binary probit choice model as the first step of a Type II Tobit model (Heckman, 1979). The probit regression models the propensity of customers to use our Facebook application (and hence to be included in the study) and provides a correction factor for self-selection to include in the engagement model. The regression is defined as a linear function of both transactional data and demographics (see e.g., Kumar et al., 2016):

$$Application\ usage_i^* = \beta_{00} + \beta_{01}\, Recency_i + \beta_{02}\, Customer\_tenure_i + \beta_{03}\, Gender_i + \beta_{04}\, Age_i + \varepsilon_i\,. \quad (2)$$

We expect younger men to use the app more often, given high digital awareness among young people and the relatively masculine soccer culture. Further, we expect higher team involvement in terms of recency and customer tenure, to lead to a higher probability to use the app. We assume a customer uses the app when the latent app usage variable, $Application\ usage_i^*$, is larger than zero. We do not observe this latent variable however, and only observe a binary variable $Application\ usage_i$ indicating actual app usage. Hence, we map the latent usage to the binary variable as follows: $Application\ usage_i = 1$ if $Application\ usage_i^* > 0$ and $Application\ usage_i = 0$ if $Application\ usage_i^* \leq 0$. Subsequently, we can derive the Inverse Mills Ratio (IMR) from the probit regression as follows:

$$\lambda = \varphi(\beta X)/\Phi(\beta X)\,, \quad (3)$$

where λ as usual indicates the IMR, and φ and Φ indicate the probability and cumulative density functions, respectively. The IMR is a monotone decreasing function of the probability that an individual self-selects into the sample. The engagement model can be seen as the second step

of the Type II Tobit model, which is dependent on the selection equation. By incorporating the IMR into the engagement model equations as an explanatory variable, we correct for potential endogeneity issues resulting from self-selection. If the IMR-coefficient in the engagement model is significant, this indicates that self-selection is indeed an issue.

Engagement Model Specification. We model engagement following the choice-then-quantity approach of (Kumar et al., 2008), considering the non-contractual nature of our context. This model follows the always-a-share approach to measuring customer lifetime value (CLV or direct engagement), assuming customers never terminate their relationship with a firm but may have periods of dormancy. Thus, a customer may return after some non-purchase period. A customer's (direct) engagement is defined as the net present value of his/her future cash flows, calculated over a three-year period as is common in CLV calculations (Kumar et al., 2008). Following (Kumar et al., 2008), engagement is measured as:

$$Engagement_i = \sum_{t=T+1}^{T+3} \frac{p(Purchase_{it} = 1) * \widehat{CM_{it}}}{(1+r)^{t-T}} - \frac{\overline{MC} * \widehat{MT_{it}}}{(1+r)^{t-T}} \quad , \qquad (4)$$

where,

$Engagement_i$ = (direct) Engagement for customer *i*,

p($Purchase_{it}$) = Predicted probability of purchase for customer *i* in year *t*

$\widehat{CM_{it}}$ = Predicted contribution margin of customer *i* in period *t*

$\overline{MC}$ = Average cost for a single communication (e-mail in this context; this is estimated to be € 0.89 by the soccer team)

$\widehat{MT_{it}}$ = Predicted marketing contacts (e-mails) for customer *i* in year *t*

t = year index

T = Marks the end of the observation phase, and

R = yearly discount rate (0.15 as is common in CLV studies, e.g. (Kumar et al., 2008))

The engagement formula thus consists of all revenue a customer brings minus the costs of marketing actions. While it is common to model the number of marketing contacts endogenously (e.g., Kumar et al. 2008), this is not the team's current practice. Marketing contacts, in this case (undirected) e-mails, are sent to every customer who provided his/her e-mail address. In order to verify this statement, we divide customers into three spending groups: low (e.g., student rate), average, and high (e.g., VIP). For each group, we select those who agree to receive e-mails and calculate the average number of e-mails sent to them. The results reveal no differences across groups. This is confirmed by an ANOVA (F-statistic= 0.4881, p =

0.4849), which shows there is no need to account for endogeneity. Moreover, we investigated whether there were systemic differences in marketing contacts over time. It appears that for the last year in our sample, the number of e-mails sent was higher compared to the previous years, which is attributed to a change in marketing agency contracted by the team. However, also in this last year, there are no differences across the different groups. Taking these results into account, we further model the predicted marketing contacts in a constant way as there are no specific drivers for sending communications ($MT_{i0} = MT_{i1} = MT_{i2} = \ldots$), but we do include an interaction effect between the last year in our analysis (2014) and the contact volume by the firm to account for the time specific changes.

The engagement formula requires two other concepts to be predicted: (1) purchase probability for customer *i* in year *t* and (2) contribution margin of customer *i* in year *t*. Contribution margin can only be observed if customers purchase. Hence, we use a Type II Tobit specification to obviate potential selection bias when measuring contribution margin. In modeling the purchase equation, we assume customer *i* will buy only when the latent utility is higher than zero. However, we do not observe this latent utility; only the buy vs. no-buy decision. We map the latent utility to this decision using a binary probit choice model:

$$Purchase_{i,t}^{*} > 0 \; if \; Purchase_{i,t} = 1 \qquad (5)$$

$$Purchase_{i,t}^{*} \leq 0 \; if \; Purchase_{i,t} = 0 \, .$$

Then, we model the latent utility $Purchase_{i,t}^{*}$ as a linear function of the predictor variables:

$$Purchase_{i,t}^{*} = \beta_{1i} + \beta_1 \, x_{1i,t} + u_{1i,t} \, , \qquad (6)$$

where β_{1i} is a vector of customer specific intercepts, β_1 is a vector of coefficients, $x_{1i,t}$ is a vector containing predictor variables and $u_{1i,t}$ captures the error term. Similarly, we assume the latent variable CM_{it}^{*} to represent the amount of purchases of customer *i* in period *t*:

$$PurchaseAmount_{it}^{*} = \beta_{2i} + \beta_2 \, x_{2i,t} + u_{2i,t} \, , \qquad (7)$$

where β_{2i} is again a vector of customer specific intercepts, β_2 is a vector of coefficients, $x_{2i,t}$ is a vector containing predictor variables for the contribution margin equation and $u_{1i,t}$ captures the error term. The latent contribution is observed when a customer purchases:

$$PurchaseAmount_{it} = PurchaseAmount_{it}^{*} \qquad (8)$$

$$if \; Purchase_{i,t} = 1, \text{ otherwise unobserved.}$$

Our dataset consists of panel data, considering several subsequent purchases over time as specified in the purchase incidence and contribution margin equations. Hence, we cannot apply the simple selection model, instead using the random-effects variant (Bruce et al., 2005; Verbeek and Nijman, 1992), as executed in Limdep 11. Parameters α_{1i} and α_{2i} represent random effects (instead of simple intercepts), assumed to be bivariate normally distributed with zero means, standard deviations σ_1 and σ_2 and correlation θ. We specify the random-effects variant as selectivity comes from two sources, i.e., the correlation of the error $u_{1i,t}$ and $u_{2i,t}$ and of the random effects α_{1i} and α_{2i} (Greene, 2016). We jointly fit purchase incidence and contribution margin equations via maximum simulated likelihood instead of a two-step approach in Limdep, which implies no IMR variable for this selection bias.

We use two broad categories of variables. First, we include demographics and control variables capturing aspects of customer-team interactions (Kumar et al., 2008). The buying equation includes a lagged purchase indicator (*Purchase*$_{t-1}$), lagged average contribution margin (*Paid Price*), customer tenure (*Tenure*), gender (*Gender*), the number of messages (e-mails) sent to the customer (*Contact Volume*), and email click-rate (*Click-Through Rate*) as an indicator of interest in team communications. The contribution margin equation includes the lagged contribution margin (*Purchase Amount*$_{t-1}$) and lagged average contribution margin. We also add the lagged percentage of home matches attended to both equations (*Consumption*). All customers included bought a season ticket; those not attending a high percentage of matches may consider their money (partially) wasted and may be less likely to buy a season ticket the following year, or may select a lower-priced ticket.

Second, we include our main variables of interest. First, we include lagged predicted customer sentiment (*Customer*$\widehat{Sentiment}$), operationalized as the average predicted customer sentiment per season, i.e., averaged over all matches per customer, representing the customer's average experience with the team during the past season. As mentioned, this variable is the link between the customer sentiment and engagement models. Next, we include whether or not a customer has liked the team's Facebook fan page (*Page like*). We expect higher average (online) customer sentiment as well as a fan page like to result in a higher probability of buying and a higher contribution margin, since these can be considered forms of engagement (Goh et al., 2013; Kumar et al., 2016; Rishika et al., 2013). Moreover, SM UGC increases customer–firm identification, which in turn, increases willingness to pay (Homburg et al., 2009). Finally, we include the moderator variable *Share of engagement*, operationalized as a user's number of (online) interests, to account for his/her online activity and number of distinct interests. We

expect a higher number of interests (lower share of CE) to result in a lower direct CE. The final equations have the following form:

$$Purchase^{*}_{i,t} = \beta_{10} + \beta_{11,i} + \beta_{12}\, \widehat{CustomerSentiment}_{i,t-1} + \beta_{13} Share\ of\ Engagement_{i,t-1} + \beta_{14} Share\ of\ Engagement_{i,t-1} * \widehat{CustomerSentiment}_{i,t-1} + \beta_{15}\, PageLike_{i,t-1} + \beta_{16}\theta_t + \beta_{17}\, Purchase_{i,t-1} + \beta_{18}\, PricePaid_{i,t-1} + \beta_{19}\, Tenure_{i,t-1} + \beta_{110}\, ContactVolume_{i,t-1} + \beta_{111}\, ContactVolume_{i,t-1} * Year2014 + \beta_{112}\, ClickThroughRate_{i,t-1} + \beta_{113}\, Consumption_{i,t-1} + \beta_{114} Gender_{i,t-1} + \beta_{115}\, IMR_i + u_{1i,t}\,, \quad (9)$$

$$PurchaseAmount^{*}_{i,t} = \beta_{20} + \beta_{21,i} + \beta_{22}\, \widehat{CustomerSentiment}_{i,t-1} + \beta_{23} Share\ of\ Engagement_{i,t-1} + \beta_{24} Share\ of\ Engagement_{i,t-1} * \widehat{CustomerSentiment}_{i,t-1} + \beta_{25}\, PageLike_{i,t-1} + \beta_{26}\theta_t + \beta_{27}\, PurchaseAmount_{i,t-1} + \beta_{28}\, Tenure_{i,t-1} + \beta_{29}\, ContactVolume_{i,t-1} + \beta_{210}\, ContactVolume_{i,t-1} * Year2014 + \beta_{211}\, ClickThroughRate_{i,t-1} + \beta_{212}\, Consumption_{i,t-1} + \beta_{213} Gender_{i,t-1} + \beta_{214}\, IMR_i + u_{1i,t}\,, \quad (10)$$

The second term in each equation represents the customer specific intercept and the variable θ_t represents a vector of year dummies, accounting for team and external factors that might vary by year. Thus, the intercept is both customer-specific and time varying. The IMR included in both equations represents the Inverse Mills Ratio calculated from the selection equation. There is no second IMR factor in the contribution margin equation, which would come from the selection based on purchase incidence, as these equations are jointly estimated by maximum simulated likelihood. This estimation does not use a two-step method and hence does not create or use an IMR variable. Appendix B provides a detailed overview of the estimation procedure. A list of variables used for the engagement model is given in Table 3.3, and descriptive measures of this engagement model, distributions of both purchase incidence and contribution margin, and correlation matrices are given in Appendix C.

Table 3.3: Overview of the variables for Engagement modeling

Variable	Description		
App Usage Equation			
Recency $_{t-1}$	Recency of the last purchase of a season ticket amount (at time t-1)		
Tenure $_{t-1}$	Length of relationship of the focal customer with the company at time *t*-1		
Gender	Gender of the focal customer		
Age	Age of the focal customer		
Engagement models			
Dependent Variables	*Measures of direct engagement*	*Purchase incidence*	*Contribution margin*
$Purchase_t$	Dummy indicating whether a season ticket was bought at time *t*	X	
$PurchaseAmount_t$	Amount spent on season tickets at time *t*		X
Variables of interest	*Measures of online engagement*		
$\widehat{Customer\ Sentiment}_{t-1}$	Predicted customer sentiment during period *t*-1	X	X
Share of Engagement $_{t-1}$	Number of different categories the customer has liked on Facebook during *t*-1	X	X
$\widehat{\text{Customer Sentiment}}_{t-1}$ * *Share of Engagement* $_{t-1}$	Interaction effect between predicted customer sentiment and share of engagement in period *t*-1	X	X
PageLike $_{t-1}$	Dummy indicating whether the customer liked the company's Facebook brand page in *t*-1	X	X
Control variables	*Control variables for the engagement models*		
ϑ_t	Dummy variables indicating the period *t*	X	X
Tenure $_{t-1}$	Length of relationship of the focal customer with the company at time *t*-1	X	X
$Purchase_{t-1}$	Dummy indicating whether a season ticket was bought at time *t*-1	X	
$PurchaseAmount_{t-1}$	Amount spent on season tickets at time *t*-1		X
PricePaid $_{t-1}$	Average amount spent on season tickets by the focal customer at time *t*-1	X	
ContactVolume $_{t-1}$	Number of email messages sent by the company to the focal customer during period *t*-1 (logarithm)	X	X
ContactVolume $_{t-1}$ * *Year2014*	Interaction effect of contact volume and the year 2014 (to account for change in marketing agency)	X	X
Click-through Rate $_{t-1}$	Click through rate of the customer on emails received by the company during period *t*-1	X	X
$Consumption_{t-1}$	The percentage of total (home) matches attended by the customer during period *t*-1	X	X
Gender	Gender of the focal customer	X	X

6. Results

6.1. Customer Sentiment

The results for customer sentiment are shown in Table 3.4 (year dummies are not included as they are not relevant for interpretation). The results indicate that the customer sentiment analysis model, based on the valence of comments per user and match, has adequate fit over a null random model (with random components per user and match), with a likelihood-ratio $\chi^2(22) = 1023.3$, $p < 0.001$.

The parameter estimates for CX, based on the objective measures of performance, show that these variables have the expected signs, but that only a part of them were significant. We note that all were significant before including match-specific intercepts (results not shown), and that these intercepts account for most of the variance in these parameters. As expected, a win (loss) results in higher (lower) customer sentiment compared to a draw (α_3 = 1.008 for wins, = -0.353 for losses; both $p < 0.01$). The number of red and yellow cards (α_4 and α_5) are negative, as expected, but not significant. Further, with regard to the type of match, only a European match results in significantly more positive customer sentiment than a cup match (α_6 =0.299, $p < 0.05$). Finally, the effect of opponent quality (α_7) is insignificant.

Next, we look at the variables related to firm generated content. The number of firm posts on its Facebook page (i.e., MGC) has a positive impact on customer sentiment (α_8 = 0.299, $p < 0.01$). The interaction effect between result of the match and MGC enables us to test the moderator effect (α_9 = 0.284 for losses, -0.290 for wins; both $p < 0.01$). We see that both the interaction effects of MGC with wins and losses (compared to draws) are significant. Figure 3.3 shows us the interaction plot, which allows us to better understand this effect.

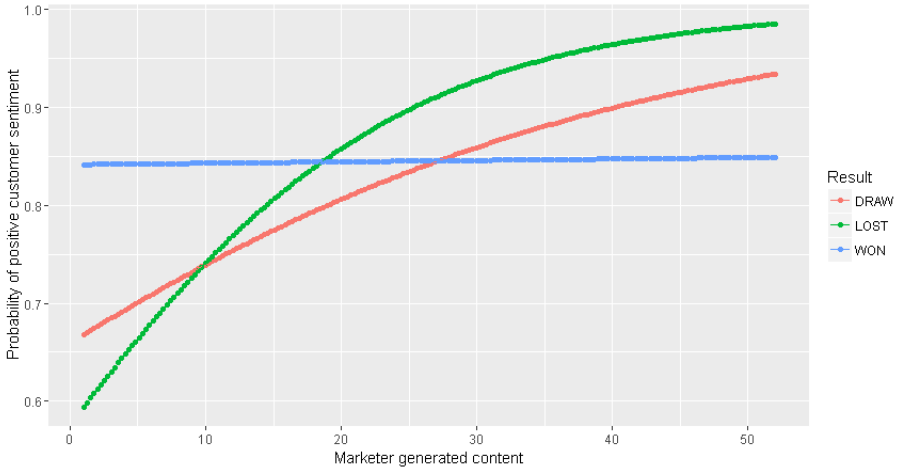


Figure 3.3: Interaction plot between MGC and match result

The plot, with MGC volume on the x-axis and customer sentiment on the y-axis, shows that overall, there is a positive relationship between the number of positive posts and customer sentiment. However, this effect is much more pronounced for draws and especially losses, even so much that with a higher number of posts, the customer sentiment for losses and draws turns out to be higher than for wins. This could indicate that in neutral and negative CX encounters, in our case represented by draws and losses, the impact of MGC is more important compared to explicit positive cases. We will further explore this in the discussion section by means of a counterfactual analysis.

Table 3.4: Customer sentiment equation results

Variables	Estimate	z-score
Intercept	0.819 ***	5.316
Result (Lost)$_m$	-0.353 ***	-3.924
Result (Won)$_m$	1.008 ***	12.413
RedCards$_m$	-0.060 *	-1.887
YellowCards$_m$	-0.035	-1.178
TypeMatch (Eur)$_m$	0.299 **	2.250
TypeMatch (Nor)$_m$	0.130	1.088
TypeMatch (PO)$_m$	0.234 *	1.770
QualityOpponent (Medium)$_m$	0.107	1.310
QualityOpponent (High)$_m$	0.007	0.086
MGC $_{u,c,m}$	0.299 ***	7.044
*ResultLost$_m$ * MGC $_{u,c,m}$*	0.284 ***	4.536
*ResultWon$_m$ * MGC $_{u,c,m}$*	-0.290 ***	-6.163
Home Match$_m$	-0.129 **	-2.183
Likes MGC Post$_{u,c,m}$	0.275 ***	16.743
EventFacebook$_{u,m}$	0.145 *	1.789
EventAttending $_{u,m}$	-0.062	-0.819
Customer Sentiment $_{u,c,m-1}$	0.093 ***	3.376
Other UGC Valence $_{u,c,m}$	0.095 ***	7.630
Other UGC Volume $_{u,c,m}$	0.098 ***	6.474
Comment length $_{u,c,m}$	-0.085 ***	-6.547
Log-Likelihood	-22,938.6	
AIC	45,931.2	

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized

Finally, we note that some but not all of the control variables are significant. First, the variables indicating event attendance, both on Facebook (α_{13}) and actual attendance (α_{14}), are not significant on a 5%-significance level, while a home match (α_{11} = -0.129, p < 0.05) is significant, indicating that fans might have higher expectations when playing at home. Next, we see that likes on posts of the team (α_{12} = 0.275, p < 0.05), as a measure of SM team identification, are indeed related to more positive customer sentiment. Moreover, we see that customer sentiment, as mentioned by Verhoef et al. (2009), is also affected by previous sentiment (α_{15}= 0.093, p < 0.01). We further see that user generated content by others is influential, both in volume (α_{17}=0.098, p<0.01) as in valence (α_{16}=0.095, p<0.01). The results related to others' UGC valence replicates the findings of Homburg, Ehm and Artz (2015). However, while the results related to others' UGC volume are not in line with the findings of, for instance Moe and Trusov (2011), we note that these authors investigate product ratings rather than customer sentiment. Finally, consistent with Homburg, Ehm and Artz (2015), we find that a longer comment text in general indicates lower sentiment (α_{18}= -0.085, p<0.01).

6.2. Engagement

Table 3.5 presents results of the selection equation. The overall model is significant (likelihood-ratio $\chi^2(4) = 1332.3;\ p < 0.01$) as are all parameter estimates. This indicates that customers indeed may be self-selected into the sample. Moreover, the signs of the parameter estimates are as expected, indicating face validity. A higher age results in a lower probability to use the application, which is plausible given the higher digital awareness among younger people. Second, male customers and customers with longer tenure also have higher application usage probabilities. Together with the significant IMR in the purchase incidence and contribution margin equations, this validates the need to accommodate selection bias.

Table 3.5: Application usage selection equation

Variables	Estimate	z-score
Intercept	-0.735 ***	-81.610
Recency	0.018 *	1.842
Tenure	0.069 ***	6.755
Gender	0.069 **	2.054
Age	-0.363 ***	-33.685

Note: * p<0.1, ** p<0.05, *** p<0.01 ; coefficients are standardized

The results of the engagement model are shown in Table 3.6 (year dummies are not included as they are not relevant for interpretation). As the models are jointly estimated, only one log-likelihood and AIC value is available per joint buying and contribution margin model. These values are shown below Table 3.6.

The information criterion in itself does not provide us with a large amount of information. However, comparing these results to the ones in Appendix A without customer specific intercepts indicates that the AICs are lower when random intercepts are included. We thus confirm that including customer heterogeneity is crucial when analyzing engagement and experience (Pansari and Kumar, 2017). The time-varying intercepts, not shown in the results, are all significant, indicating that this may be necessary to absorb season-specific shocks.

Most control variables are significant for both purchase incidence and contribution margin equations. We first focus on the purchase incidence equation. Prior research reports positive impacts of previous purchase behavior and price on purchase probability, both of which are confirmed (β_{17} = 1.07 and β_{18} = 0.09 respectively; both $p < 0.01$). Tenure (β_{19}) however, is not significant. Prior research also suggests marketing communications are positively related to purchase probability, which is only partially confirmed: contact volume (β_{110} = -0.05, $p < 0.01$) is negatively related to purchase incidence, but with a significant, positive interaction effect in 2014 (β_{111} = 0.07, $p < 0.01$). Click-through rate is positive and significant, as expected (β_{112} = 0.06, $p < 0.01$). We confirm that the percentage of matches attended is positively related to purchase incidence (β_{113} = 0.19, $p < 0.01$). Finally, gender is significant; males have a higher purchase propensity (β_{114} = 0.06, $p < 0.05$).

With regard to the contribution margin equation, we note that all control variables are significant and have the expected sign, except for the contact volume which is significant and negatively related to the purchase amount (β_{29} = -2.72, $p < 0.05$) but also has a positive significant interaction effect in 2014 (β_{210} = 3.72, $p < 0.05$). The purchase amount spent on season tickets last year is by far the most important predictor.

With regard to our main variables of interest related to SM, we see that the variables have their expected signs in both equations, i.e. both (predicted) customer sentiment (β_{12} = 0.04 and β_{22} = 1.67) and Facebook fan page likes (β_{15} = 0.02 and β_{25} = 1.81) are positively related to direct CE, while a higher number of online interests (and hence a lower share of engagement; β_{13} = 0.00 and β_{23} = -0.56) relate to lower direct CE. However, our results show that only the predicted customer sentiment is significant in modeling purchase incidence ($p <$

Table 3.6: Engagement model results

	Purchase incidence		Contribution margin	
Variables	Estimate	z-score	Estimate	z-score
Intercept	-0.28 ***	-6.77	165.66 ***	64.94
$\widehat{Customer\ Sentiment}$	0.04 ***	4.46	1.67 **	2.03
Share of Engagement	0.00	-0.33	-0.56	-0.67
$\widehat{Customer\ Sentiment}$ Share of Engagement*	0.00	0.26	0.41	0.69
Page Like	0.02	0.86	1.81	1.12
$Purchase_{t-1}$	1.07 ***	31.60		
$PurchaseAmount_{t-1}$			75.66 ***	303.23
Price paid	0.09 ***	5.47		
Tenure	0.01	1.21	5.63 ***	6.65
Contact Volume	-0.05 ***	-3.52	-2.72 **	-2.21
*Year2014*Contact Volume*	0.07 ***	3.21	3.72 **	2.21
Click-through Rate	0.06 ***	5.81	3.68 ***	4.50
Consumption	0.19 ***	12.54	2.57 **	2.41
Gender	0.06 **	2.46	8.93 ***	5.31
IMR	0.02 **	2.42	14.58 ***	22.85
σ		0.01	0.06	
ρ		0.89 ***	317.09	
Year dummies		*Included*		
AIC		198,046.9		
Log-Likelihood		-98,986.43		

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized; subscripts are not included for clarity, except for the lagged dependent variables

0.01) and contribution margin ($p < 0.05$), and not Facebook fan page likes nor the share of engagement. Thus, we can state that customer sentiment is relevant for modeling engagement over and above our control variables, and also more relevant compared to Facebook fan page likes, the most commonly used measure in literature. Finally, we can state that the share of engagement, measured by the number of online like categories, does not influence direct CE.

We further note that the IMR is significant for both the purchase incidence and contribution margin equations, indicating that self-selection was indeed an issue. A higher IMR-value indicates a lower probability to use the application. Given the positive sign of the IMR parameter coefficients (β_{115} = 0.02 and β_{214} = 14.58), we can conclude that the lower the probability to use the application, the higher the propensity to purchase and the higher the average contribution margin. This is, however, not surprising since the sample consists mainly of younger people, with lower season ticket fares (average season ticket price for sample customers is €147 vs. €155 for out-of-sample customers). The logical interpretation of the IMR coefficients adds face validity to the results.

Finally, σ and ρ represent the standard error for the contribution margin equation and the correlation between the residuals of the contribution margin and purchase incidence equation, respectively. These parameters are used in the estimation of the Type II Tobit model for the two equations. The high correlation factor (ρ) indicates the need to use a Type II Tobit specification for modeling contribution margin and purchase incidence.

6.3. Robustness checks

We had to make choices in building the customer sentiment and engagement models. To check the robustness of our choices and findings, we estimate variants of both models. First, we re-estimated the customer sentiment model using a different time-window for the collection of comments. Instead of using a 2-day window starting from the end of the match, we use a 1 and 3-day time window. Note that we cannot use a longer time window, since some matches are only four days apart. In general, the results (see appendix E) of these models were similar to the ones presented previously. Second, the presented analyses for the engagement models use a regular average of the predicted customer sentiment over the entire season to arrive at an average prediction. However, because matches (and comments) at the end of the season may be more influential compared to matches earlier in the season, we estimate weighted models with more weight given to more recent matches. The results (appendix F) indicate no changes in the overall conclusions. Finally, we include network effects for a sample of our data, and the

presence of network data does not affect our results (see appendix G for a discussion of the identification issue with network data, and appendix H for the results; see also the discussion section for further elaboration on these results).

7. Discussion

In this paper we set out to study the potential for MGC on SM to amplify or temper the influence of objective CX encounters on direct CE (as measured by purchase incidence and contribution margin). First, we built a database in which the team's internal customer data were matched with a comprehensive set of variables taken from customers' Facebook profiles, over a period of four years. Second, using comments posted on the organization's Facebook page by individual customers within each of the 212 match windows, we estimated each customer's sentiment per season over the time period of our study, the goal of which was to then use these estimates to model direct CE. Third, we estimated each customer's direct engagement using both the sentiment estimates and a range of both SM variables and variables from the firm's internal database.

This approach enables us to answer our research questions. First, we demonstrate that marketers can influence customer sentiment related to identifiable experiences via their SM contributions, in our case Facebook posts. This is in line with previous research stating that MGC can positively influence customer sentiment (e.g., Colicev et al., 2018; Homburg et al., 2015). However, our study is unique in showing that these results hold when evaluating sentiment related to specific CX encounters, taking into account the objective performance of these encounters. This is therefore the first study to show that MGC can improve customer sentiment when CX performance is suboptimal (a draw or a loss in our case) even to the point where it surpasses the level of positive sentiment under optimal conditions (a win). Second, we show that customer sentiment is effective in modeling both purchase likelihood and contribution margin components of direct CE; higher sentiment leads to a higher expected CE. While this is consistent with previous literature on both individual and firm-level outcome measures (Colicev et al., 2018; Goh et al., 2013; Hewett et al., 2016; Kumar et al., 2016; Xie and Lee, 2015), our customer sentiment metric is linked to actual and identifiable experiences. Third, because 'liking' is the most studied SM variable (also see Goh et al., 2013; Kumar et al., 2016; Mochon et al., 2017; Rishika et al., 2013; Zhang and Pennacchiotti, 2013) we also add this variable to our model and investigate its relative importance in predicting direct CE. In contrast to previous research, we find that page likes are not significantly related to purchase

likelihood (Rishika et al., 2013) or contribution margin. Previous research reports mixed results with regard to the added value of page likes for revenue streams. For instance, Goh et al. (2013) and Kumar et al. (2016) find positive influences, while John et al. (2017) and Xie and Lee (2015) do not. Further analysis (Appendix H[2]) reveals that page likes are significant in models with only SM and network related variables. However, when all controls are added, page likes become insignificant while customer sentiment remains highly significant. This result may suggest one reason for mixed results in previous studies, and also underscores the need to take into account proper controls in assessing the value of SM variables.

Further, we explore another aspect of SM - network information. The results of an engagement model with network variables included, which we were able to estimate on a sample of our data for which network information was present (2386 of 4783 customers), shows that the network structure of SM data may also be valuable for modeling CE (Appendix H[2]). Variables constituted from the social network data all have the expected influence on purchase likelihood and contribution margin based on prior research, illustrating the potential value of network variables in modeling important customer- and firm-level outcomes. We use the number of friends who bought tickets the prior year (Network Customers), average homophily with friends based on age, gender and ticket seat location (Homophily) and the percentage of defectors among friends in the last season (Network Defectors) for the purchase incidence equation. In line with (Nitzan and Libai, 2011), the percentage of defectors is negatively related to purchase propensity. However, in contrast to their research, the number of customer-friends is positively related to buying incidence. An explanation may be found in the role of friends as reference groups in sports contexts. The larger the reference group, the greater the social acceptance for event participation, and the higher re-patronage intentions (Wakefield, 1995). While we expected homophily to be positive based on expected great social acceptance, we find it to be insignificant. We conclude that our study offers an additional contribution in its approach to using social networks extracted from SM to constitute network variables, and demonstrating their value beyond customer sentiment and page likes, with the latter not even reaching significance, in predicting direct CE.

[2] The results are structured as follows: the first table gives the results for purchase incidence, for a model with only SM variables, a model with SM variables and network variables and a complete model including all control variables. The second table gives the results for the contribution margin equation for these 3 models.

Next, we discuss important implications of our study. In doing so, we illustrate how managers can use our results based on a series of counterfactual analysis. We also offer directions for researchers to build on the approach introduced here.

7.1. Theoretical implications

Our results have interesting implications for marketing theory as well. First, we contribute to the growing literature on CE, and to CE theory more specifically by demonstrating potential firm influences beyond more traditional marketing activities aimed at creating awareness. Building on the CE theory framework proposed by Pansari and Kumar (2017), our results suggest MGC as a moderator on the effect of experience characteristics on both emotion and satisfaction (reflected in our measure of customer sentiment) beyond characteristics of the product (convenience), firm, and industry. Interestingly, these results might also offer direction in linking CE theory with the theory of CEM proposed by Harmeling et al. (2017). However, rather than a direct impact of firm communications on CE, as conceptualized by Harmeling et al. (2017), our conceptualization and results support its moderating impact based on actual brand experiences.

Our results also contribute to the growing literature investigating the role of SM as a source of insights related to CE. Our study is the first to show that researchers should go beyond 'liking' - the most studied SM variable in literature (see Goh et al., 2013; Kumar et al., 2016; Mochon et al., 2017; Rishika et al., 2013; Zhang and Pennacchiotti, 2013). Our results show that 'liking' a page is considerably less important for direct CE than customer sentiment, but that conclusions on its significance may be dependent on the completeness of the model used. Thus, we highlight the importance of considering a comprehensive set of relevant variables in order to understand drivers of direct CE.

Finally, our findings regarding the significant role of customers' networks might also offer direction in linking network and CE theories. From a network theory perspective, our finding that the number of customer-friends (defectors) is positively (negatively) related to buying incidence might suggest the importance of social influence, which occurs when a an individual varies his or her behavior based on that of others in a social system (Leenders, 2002). Such influence can be based on information originating from others in a network (Robins et al., 2001), such as digital content shared in SM. From a network theory perspective, the greater degree centrality for a given customer, such that the customer is connected to a greater number of others, the greater the potential for such influence. Considering these findings in conjunction

with ours suggests the importance of these network characteristics for a host of important firm- and customer-level outcomes. Customers' social value, or the monetary value created by their social interactions (Kumar et al., 2010a; Libai et al., 2013), may thus extend to their influence on others' direct CE. In addition, the density of a firm's own network in terms of its connections with customers may enhance the ability of its communications to lift CE based on customers' brand experiences.

7.2. Counterfactual Analyses

In this section we aim to offer direction for managers in making resource allocations in their efforts to improve CE. More specifically we aim to provide insight into what it takes to influence CE above and beyond objective performance and their related experiences, the most important factor, according to our analyses, in driving sentiment and ultimately engagement. In the first counterfactual analysis we aim to compute the boundary cases of objective performance, which in our case can be done by simulating all matches to be wins, losses or draws. Because these are boundary cases, this will allow us to deduce the maximum impact of objective performance. We compare this with the current actual data. In the second counterfactual analysis, we compute the influence of MGC by varying the level of MGC (described in further detail below). This will allow us to compare the impact of objective CX performance with that of MGC. In the final simulation analysis, we combine different levels of both objective performance and MGC to deduce if MGC can amplify or temper the impact of the experience encounter. For all simulation analyses, we consider the absolute value (and changes) of mean purchase propensity, mean contribution margin and overall direct CE (which is the sum of direct engagement over all customers, and sometimes called customer equity, e.g., Rust et al., 2004).

Table 3.7 shows the results of simulation analyses. In analysis 1 we simulated the boundary cases of the performance (at mean values of predicted customer sentiment) and compared it to the actual observed situation. We find that going from the actual match result to the most positive performance (assuming all "wins") would result in an increase of 1.94% in customer equity. In other words, the firm is losing 1.94% in potential CLV across all customers due to more negative performances. The more neutral and negative performances ('all draws' and 'all losses') result in worse customer equity than the actual scenario.

In analysis 2, we simulate a variety of MGC percentage increases. Since the variables are standardized, these percentage values refer to the percentage of the standard deviation of that

variable. So, in this case, 50% refers to 50% of one standard deviation of MGC. Adopting a MGC posting policy with an increase of 50%, which equates to approximately four more messages per match on average, results in a 0.49% increase in customer equity. Increasing MGC by 100% results in a 0.97% increase in customer equity, or about half that of the most positive objective performance scenario of 1.94%. However, the effort needed to reach a perfect win rate is arguably a different order of magnitude than merely increasing the number of posts. For example, replacing 30% of the more negative performances (draws and losses) by positive performances, which would be very good results, would result in an increase of only 0.56% (not shown), comparable to the result based on a 50% increase in MGC. Note that the percentage increase in MGC has diminishing returns. Specifically, adding more MGC at already higher levels does not yield the same proportional increase in customer equity as adding MGC at lower levels. This is not entirely clear from Table 3.7, since we are still at relatively low values of MGC (e.g., average MGC for losses is 3.64 posts within the match window, and one standard deviation is approximately 8 posts). The interaction plot (Figure 3.3) shows that at this range, the lines are fairly proportional, but that they tend to flatten for higher levels of MGC. This also supports previous literature on the effects of MGC in online forums (Homburg, Ehm and Artz, 2015). Finally, most of the gain in customer equity comes from the increase in purchase incidence as opposed to contribution margin, as can be seen by their respective percentage increases.

Table 3.7: Simulation analysis 1 & 2

Decision variable	Value of decision variable	Mean PI as % (% increase vs baseline)	Mean CM in $ (% increase vs baseline)	Customer Equity in $ (% increase vs baseline)
	Baseline (actual)	63.2	206.13	4,261,905
Match result	All wins	64.2 (1.58)	207.38 (0.61)	4,344,615 (1.94)
	All draws	62.5 (-1.11)	205.14 (-0.48)	4,202,072 (-1.40)
	All losses	61.8 (-2.22)	204.22 (-0.93)	4,143,576 (-2.78)
MGC	10% increase (in sd)	63.3 (0.08)	206.19 (0.03)	4,266,057 (0.10)
	50% increase (in sd)	63.5 (0.39)	206.44 (0.15)	4,282,675 (0.49)
	90% increase (in sd)	63.6 (0.69)	206.69 (0.27)	4,299,094 (0.87)
	100% increase (in sd)	63.7 (0.77)	206.75 (0.30)	4,303,134 (0.97)

Table 3.8 contains the results of simulation analysis 3. In a scenario with a perfect win rate scenario and an increase in MGC from the mean value by 50% (100%) of the standard deviation,

there would be a 0.02% (0.03%) increase in customer equity. Thus, when the objective performance criteria are perfect and CX is good, MGC level does not have a large influence on customer value. In the "all draw" scenario, an increase in MGC from the mean value by 50% (100%) would result in a 0.68% (1.35%) increase in customer equity. These numbers become more favorable in the "all loss" scenario such that increasing MGC from the mean level by 50% (100%) results in a 2.45% (2.72%) increase in customer equity. In sum, based on the challenges inherent in improving performance, such as accounting for a wide range of uncontrollable factors, these results might also suggest that managers may find a greater return from increasing their investments in MGC as opposed to focusing exclusively on improvements in performance during individual experience encounters. In addition, increasing MGC may be most effective in neutral or negative encounters.

Table 3.8: Simulation analysis 3

Match Result	MGC level	Mean PI as % (% increase vs match result baseline)	Mean CM in € (% increase vs match result baseline)	Customer Equity in $ (% increase vs match result baseline)
All wins	Mean-level for wins	64.2	207.38	4,344,615
	50% increase (in sd)	64.2 (0.01)	207.39 (0.01)	4,345,299 (0.02)
	100% increase (in sd)	64.2 (0.03)	207.40 (0.01)	4,345,981 (0.03)
All draws	Mean-level for draws	62.5	205.14	4,202,072
	50% increase (in sd)	62.8 (0.56)	205.60 (0.22)	4,230,813 (0.68)
	100% increase (in sd)	63.2 (1.11)	206.04 (0.44)	4,258,855 (1.35)
All losses	Mean-level for losses	61.8	204,22	4,143,576
	50% increase (in sd)	62.5 (1.14)	205.12 (0.44)	4,200,664 (1.38)
	100% increase (in sd)	63.2 (2.23)	205.99 (0.87)	4,256,120 (2.72)

7.3. Managerial Implications

We can draw a number of important, managerially relevant conclusions from our research. First, we demonstrate the value of SM as an effective customer (sentiment) tracking mechanism, which can help managers gain insight regarding customers' experiences (de Vries et al., 2017). We demonstrate the utility of customer-level SM data for modeling behavior. Customers increasingly take to SM to provide feedback and interact with brands, as indicated by vibrant online communities. Feedback can be accessed anytime, from anywhere (with Internet access),

from any device, enabling communication immediately after experiences. Thus, we expect its viability as a resource for insights to only expand.

Second, we show that Facebook page likes may not be the "holy grail" for marketers in terms of direct CE. On the one hand, one might expect that 'liking' a brand on Facebook is positively related to engagement (CLV) because the very act of 'liking' the page results in participation in a brand's online community; on the other hand, it might be unrealistic to expect this positive relationship given that some Facebook users like hundreds of brands (John et al. 2017). From our research, we conclude that the latter is more likely to be true.

Finally, our results based on more restricted models in Appendix H show that SM information in the form of customer sentiment and page likes can be indicative of future behavior, even in the absence of behavioral data. That is, we can model purchase propensity and project direct CE of prospects for whom no behavioral data is yet observable. Thus, our results contribute to literature focused on prioritizing prospects based on their estimated CLV (Kumar and Petersen, 2012) and suggest that prospects' social networks may also represent opportunities for direct CE estimation and targeting. Coupled with findings that customers acquired based on WOM add nearly twice as much long-term value to the firm than those acquired via traditional promotional marketing tactics (Villanueva et al., 2008), these results suggest the potential power of SM networks themselves as vehicles for marketing campaigns, e.g., initiatives fostering WOM in prospects' friend network. The ability to refine such campaigns based on estimates of prospects' referral values (Kumar et al., 2010b) could further boost the potential impact on firm value.

8. Limitations and Future Research Directions

This research represents one of the few empirical demonstrations of the link between CX encounters, MGC, customer sentiment and direct CE, an issue of great interest to managers and academic researchers. However, we must acknowledge several limitations that should be considered in evaluating our findings and that may encourage future research efforts. First, our study was limited to the professional sports context. While we argue that this context is ideal for examining the phenomena under study, additional research should extend the proposed empirical analysis to other contexts. The magnitude of the effects may depend on firm specific factors, such as industry, and customer involvement needed. Nonetheless, we believe that demonstrating the positive impact of MGC on customer sentiment and its subsequent influence on CE are important findings. Our approach could be extended into other settings in which

customers have a substantial SM presence. An interesting extension would be to assess the ability of our approach in a contractual setting, in which buyers have less discretion in terms of future purchase decisions.

Our study was also limited to a time frame of four years. Although four years does enable us to observe multiple purchase opportunities and decisions, a longer window of time may yield additional insights. For example, time-varying coefficients could be used to assess the variance of the impact of customer sentiment on engagement over time.

Finally, the use of SM data from Facebook alone may be viewed as a potential limitation. While we argue that it is particularly appropriate for our context based on evidence that sports fans are particularly likely to leverage Facebook to discuss sports, there remains the possibility that insights from other platforms may be valuable and even supplement those from Facebook in helping firms assess customer sentiment and model engagement. Future studies might assess the ability of other, or a complementary set of social networking sites, to enable customer sentiment measurement and prediction, and CE modeling.

9. References

Anderson, E.W., Mittal, V., 2000. Strengthening the Satisfaction-Profit Chain. Journal of Service Research 3, 107–120.

Babić Rosario, A., Sotgiu, F., De Valck, K., Bijmolt, T.H.A., 2016. The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. Journal of Marketing Research 53, 297–318.

Baker, A.M., Donthu, N., Kumar, V., 2016. Investigating How Word-of-Mouth Conversations About Brands Influence Purchase and Retransmission Intentions. Journal of Marketing Research 53, 225–239.

Beukeboom, C.J., Kerkhof, P., de Vries, M., 2015. Does a Virtual Like Cause Actual Liking? How Following a Brand's Facebook Updates Enhances Brand Evaluations and Purchase Intention. Journal of Interactive Marketing 32, 26–36.

Bolton, R.N., 1998. A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. Marketing Science 17, 45–65.

Branscombe, N.R., Wann, D.L., 1992. Role of Identification with a Group, Arousal, Categorization Processes, and Self-Esteem in Sports Spectator Aggression. Human Relations 45, 1013–1033.

Brodie, R.J., Ilic, A., Juric, B., Hollebeek, L., 2013. Consumer engagement in a virtual brand community: An exploratory analysis. Journal of Business Research, (1)Thought leadership in brand management(2)Health Marketing 66, 105–114.

Bruce, N., Desai, P.S., Staelin, R., 2005. The Better They Are, the More They Give: Trade Promotions of Consumer Durables. Journal of Marketing Research 42, 54–66.

Bruce, N.I., Murthi, B. p. s., Rao, R.C., 2017. A Dynamic Model for Digital Advertising: The Effects of Creative Format, Message Content, and Targeting on Engagement. Journal of Marketing Research 54, 202–218.

Caruso-Cabrera, J., Golden, M., 2016. Why Marriott looks at everything you post on social media from your room [WWW Document]. URL http://www.cnbc.com/2016/08/02/why-marriott-looks-at-what-you-post-on-social-media-from-your-room.html (accessed 7.28.17).

Castellano, J., Casamichana, D., Lago, C., 2012. The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. Journal of Human Kinetics 31, 139–147.

Catalyst, 2013. Fan social media use passes a threshold [WWW Document]. Sportsbusinessdaily.com. URL http://www.sportsbusinessdaily.com/Journal/Issues/2013/09/30/Research-and-Ratings/Catalyst-social-media.aspx (accessed 7.29.17).

Chahal, H., Kaur, G., Rani, A., 2015. Exploring the Dimensions of Customer Experience and Its Impact on Word-of-Mouth: A Study of Credit Cards. Journal of Services Research; Gurgaon 15, 7–33.

Chen, Y., Fay, S., Wang, Q., 2011. The Role of Marketing in Social Media: How Online Consumer Reviews Evolve. Journal of Interactive Marketing 25, 85–94.

Chien, S.Y., Theodoulidis, B., Burton, J., 2016. Extracting Customer Intelligence by Social Media Dialog Mining: An Ontological Approach for Customer Experience Analysis. Presented at the AMA Summer Educators' Conference Proceedings, p. F-69-F-70.

Cialdini, R.B., Borden, R.J., Thorne, A., Walker, M.R., Freeman, S., Sloan, L.R., 1976. Basking in reflected glory: Three (football) field studies. Journal of Personality and Social Psychology 34, 366–375.

Clemes, M.D., Brush, G.J., Collins, M.J., 2011. Analysing the professional sport experience: A hierarchical approach. Sport Management Review 14, 370–388.

Colicev, A., Malshe, A., Pauwels, K., O'Connor, P., 2018. Improving Consumer Mindset Metrics and Shareholder Value Through Social Media: The Different Roles of Owned and Earned Media. Journal of Marketing 82, 37–56.

Cyrenne, P., 2001. A Quality-of-Play Model of a Professional Sports League. Economic Inquiry 39, 444–452.

de Vries, L., Gensler, S., Leeflang, P.S.H., 2017. Effects of Traditional Advertising and Social Messages on Brand-Building Metrics and Customer Acquisition. Journal of Marketing 81, 1–15.

Edelman, D.C., Singer, M., 2015. Competing on Customer Journeys. Harvard Business Review 93, 88–100.

Farhadloo, M., Patterson, R.A., Rolland, E., 2016. Modeling customer satisfaction from unstructured data using a Bayesian approach. Decision Support Systems 90, 1–11.

Fornell, C., Rust, R.T., Dekimpe, M.G., 2010. The Effect of Customer Satisfaction on Consumer Spending Growth. Journal of Marketing Research 47, 28–35.

Funk, D.C., 2017. Introducing a Sport Experience Design (SX) framework for sport consumer behaviour research. Sport Management Review 20, 145–158.

Godes, D., Mayzlin, D., 2004. Using Online Conversations to Study Word-of-Mouth Communication. Marketing Science 23, 545–560.

Goh, K.-Y., Heng, C.-S., Lin, Z., 2013. Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research 24, 88–107.

Greene, W.H., 2016. Sample Selection Models for Panel Data, in: Econometric Modeling Guide Limdep 11. Econometric Software, Inc.

Gupta, S., Lehmann, D.R., Stuart, J.A., 2004. Valuing Customers. Journal of Marketing Research 41, 7–18.

Harmeling, C.M., Moffett, J.W., Arnold, M.J., Carlson, B.D., 2017. Toward a theory of customer engagement marketing. J. of the Acad. Mark. Sci. 45, 312–335.

He, W., Tian, X., Chen, Y., Chong, D., 2016. Actionable Social Media Competitive Analytics For Understanding Customer Experiences. Journal of Computer Information Systems 56, 145–155.

Heckman, J.J., 1979. Sample Selection Bias as a Specification Error. Econometrica 47, 153–161.

Hewett, K., Rand, W., Rust, R.T., van Heerde, H.J., 2016. Brand Buzz in the Echoverse. Journal of Marketing 80, 1–24.

Hogan, J.E., Lemon, K.N., Libai, B., 2003. What Is the True Value of a Lost Customer? Journal of Service Research 5, 196–208.

Homburg, C., Ehm, L., Artz, M., 2015. Measuring and Managing Consumer Sentiment in an Online Community Environment. Journal of Marketing Research 52, 629–641.

Homburg, C., Koschate, N., Hoyer, W.D., 2005. Do Satisfied Customers Really Pay More? A Study of the Relationship Between Customer Satisfaction and Willingness to Pay. Journal of Marketing 69, 84–96.

Homburg, C., Wieseke, J., Hoyer, W.D., 2009. Social Identity and the Service–Profit Chain. Journal of Marketing 73, 38–54.

John, L.K., Emrich, O., Gupta, S., Norton, M.I., 2017. Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes. Journal of Marketing Research 54, 144–155.

Kelley, S.W., Turley, L.W., 2001. Consumer perceptions of service quality attributes at sporting events. Journal of Business Research, Retail Consumer Decision Processes 54, 161–166.

Kokes, A., 2017. The Integration Of Marketing And Customer Experience [WWW Document]. Forbes.com. URL https://www.forbes.com/sites/forbescommunicationscouncil/2017/12/20/the-integration-of-marketing-and-customer-experience/ (accessed 3.10.18).

KPMG, 2016. How much is customer experience worth?: Mastering the economics of the CX journey. KPMG.com.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., Kannan, P. k., 2016. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. Journal of Marketing 80, 7–25.

Kumar, V., 2018. A Theory of Customer Valuation: Concepts, Metrics, Strategy, and Implementation. Journal of Marketing 82, 1–19.

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., Tillmanns, S., 2010a. Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value. Journal of Service Research 13, 297–310.

Kumar, V., Bhaskaran, V., Mirchandani, R., Shah, M., 2013. Practice Prize Winner—Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey. Marketing Science 32, 194–212.

Kumar, V., Petersen, J.A., 2012. Statistical Methods in Customer Relationship Management. John Wiley & Sons.

Kumar, V., Petersen, J.A., Leone, R.P., 2010b. Driving Profitability by Encouraging Customer Referrals: Who, When, and How. Journal of Marketing 74, 1–17.

Kumar, V., Venkatesan, R., Bohling, T., Beckmann, D., 2008. Practice Prize Report—The Power of CLV: Managing Customer Lifetime Value at IBM. Marketing Science 27, 585–599.

Leenders, R.T.A.J., 2002. Modeling social influence through network autocorrelation: constructing the weight matrix. Social Networks 24, 21–47.

Lemon, K.N., 2016. The Art of Creating Attractive Consumer Experiences at the Right Time: Skills Marketers Will Need to Survive and Thrive. GfK Marketing Intelligence Review 8, 44–49.

Lemon, K.N., Verhoef, P.C., 2016. Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing 80, 69–96.

Libai, B., Bolton, R., Bügel, M.S., Ruyter, K. de, Götz, O., Risselada, H., Stephen, A.T., 2010. Customer-to-Customer Interactions: Broadening the Scope of Word of Mouth Research. Journal of Service Research 13, 267–282.

Libai, B., Muller, E., Peres, R., 2013. Decomposing the Value of Word-of-Mouth Seeding Programs: Acceleration Versus Expansion. Journal of Marketing Research 50, 161–176.

Luo, X., Zhang, J., Duan, W., 2012. Social Media and Firm Equity Value. Information Systems Research 24, 146–163.

Ma, L., Sun, B., Kekre, S., 2015. The Squeaky Wheel Gets the Grease—An Empirical Analysis of Customer Voice and Firm Intervention on Twitter. Marketing Science 34, 627–645.

Madrigal, R., 1995. Cognitive and affective determinants of fan satisfaction with sporting event attendance. Journal of Leisure Research; Urbana 27, 205.

Manchanda, P., Packard, G., Pattabhiramaiah, A., 2015. Social Dollars: The Economic Impact of Customer Participation in a Firm-Sponsored Online Customer Community. Marketing Science 34, 367–387.

Marketing Science Institute, 2016. "Research Priorities 2016-2018."

Micu, A., Micu, A.E., Geru, M., Lixandroiu, R.C., 2017. Analyzing user sentiment in social media: Implications for online marketing strategy. Psychol. Mark. 34, 1094–1100.

Misopoulos, F., Mitic, M., Kapoulas, A., Karapiperis, C., 2014. Uncovering customer service experiences with Twitter: the case of airline industry. Management Decision 52, 705–723.

Mochon, D., Johnson, K., Schwartz, J., Ariely, D., 2017. What Are Likes Worth? A Facebook Page Field Experiment. Journal of Marketing Research 54, 306–317.

Moe, W.W., Trusov, M., 2011. The Value of Social Dynamics in Online Product Ratings Forums. Journal of Marketing Research 48, 444–456.

Moorman, C., 2017. Capitalizing On Social Media Investments. CMOSurvey.org.

Nam, S., Manchanda, P., Chintagunta, P.K., 2010. The Effect of Signal Quality and Contiguous Word of Mouth on Customer Acquisition for a Video-on-Demand Service. Marketing Science 29, 690–700.

Ngobo, P.V., 2017. The trajectory of customer loyalty: an empirical test of Dick and Basu's loyalty framework. J. of the Acad. Mark. Sci. 45, 229–250.

Nitzan, I., Libai, B., 2011. Social Effects on Customer Retention. Journal of Marketing 75, 24–38.

Pansari, A., Kumar, V., 2017. Customer engagement: the construct, antecedents, and consequences. Journal of the Academy of Marketing Science 45, 294–311.

Pham, M.T., Goukens, C., Lehmann, D.R., Stuart, J.A., 2010. Shaping Customer Satisfaction Through Self-Awareness Cues. Journal of Marketing Research 47, 920–932.

Reinartz, W., Thomas, J.S., Kumar, V., 2005. Balancing Acquisition and Retention Resources to Maximize Customer Profitability. Journal of Marketing 69, 63–79.

Rishika, R., Kumar, A., Janakiraman, R., Bezawada, R., 2013. The Effect of Customers' Social Media Participation on Customer Visit Frequency and Profitability: An Empirical Investigation. Information Systems Research 24, 108–127.

Robins, G., Pattison, P., Elliott, P., 2001. Network models for social influence processes. Psychometrika 66, 161–189.

Rongala, A., 2016. 6 Effective Performance Metrics for Contact Center Success [WWW Document]. customerthink.com. URL http://customerthink.com/6-effective-performance-metrics-for-contact-center-success/ (accessed 3.10.18).

Rust, R.T., Lemon, K.N., Zeithaml, V.A., 2004. Return on Marketing: Using Customer Equity to Focus Marketing Strategy. Journal of Marketing 68, 109–127.

Schmitt, B., 1999. Experiential Marketing. Journal of Marketing Management 15, 53–67.

Schmitt, B.H., 2003. Customer Experience Management: A Revolutionary Approach to Connecting with Your Customers, 1 edition. ed. Wiley, New York.

Schweidel, D.A., Moe, W.W., 2014. Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. Journal of Marketing Research 51, 387–402.

Smith, A.N., Fischer, E., Yongjian, C., 2012. How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? Journal of Interactive Marketing 26, 102–113.

Sonnier, G.P., McAlister, L., Rutz, O.J., 2011. A Dynamic Model of the Effect of Online Communications on Firm Sales. Marketing Science 30, 702–716.

Statista, 2018. Social media marketing spending in the U.S. 2017 [WWW Document]. Statista.com. URL https://www.statista.com/statistics/276890/social-media-marketing-expenditure-in-the-united-states/ (accessed 3.10.18).

Stein, L., 2016. Marketers Keep Spending on Social Despite Lack of Results [WWW Document]. Advertising Age. URL http://adage.com/article/agency-news/marketers-spending-social-lack-results/302701/ (accessed 3.10.18).

Stores.org, 2017. Real-time information gives smaller retailers the upper hand [WWW Document]. STORES: NRF's Magazine. URL https://stores.org/2017/12/11/keeping-up-with-the-big-players/ (accessed 3.10.18).

Tajfel, H., Turner, J., 1979. An integrative theory of intergroup conflict, in: Hogg, M.A., Abrams, D. (Eds.), The Social Psychology in Intergroup Relations. Psychology Press, New York, NY, US, pp. 33–47.

Taparia, S., 2015. 3 Ways to Link Online and Offline Customer Experiences [WWW Document]. Chief Marketer. URL http://www.chiefmarketer.com/3-ways-linking-online-offline-customer-experiences/ (accessed 3.12.18).

Tirunillai, S., Tellis, G.J., 2012. Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance. Marketing Science 31, 198–215.

Trusov, M., Bucklin, R.E., Pauwels, K., 2009. Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site. Journal of Marketing 73, 90–102.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the social group: A self-categorization theory. Basil Blackwell, Oxford, UK: Blackwell.

van Doorn, J. van, Lemon, K.N., Mittal, V., Nass, S., Pick, D., Pirner, P., Verhoef, P.C., 2010. Customer Engagement Behavior: Theoretical Foundations and Research Directions. Journal of Service Research 13, 253–266.

Van Leeuwen, L., Quick, S., Daniel, K., 2002. The Sport Spectator Satisfaction Model: A Conceptual Framework for Understanding the Satisfaction of Spectators. Sport Management Review 5, 99–128.

Verbeek, M., Nijman, T., 1992. Testing for Selectivity Bias in Panel Data Models. International Economic Review 33, 681–703.

Verhoef, P.C., Lemon, K.N., Parasuraman, A., Roggeveen, A., Tsiros, M., Schlesinger, L.A., 2009. Customer Experience Creation: Determinants, Dynamics and Management Strategies. Journal of Retailing, Enhancing the Retail Customer Experience 85, 31–41.

Villanueva, J., Yoo, S., Hanssens, D.M., 2008. The Impact of Marketing-Induced Versus Word-of-Mouth Customer Acquisition on Customer Equity Growth. Journal of Marketing Research 45, 48–59.

Villarroel Ordenes, F., Theodoulidis, B., Burton, J., Gruber, T., Zaki, M., 2014. Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. Journal of Service Research 17, 278–295.

Voyles, B., 2007. Beyond loyalty: Meeting the Challenge of Customer Engagement. Economist, Intelligence Unit.

Wakefield, K.L., 1995. The pervasive effects of social influence on sporting event attendance. Journal of Sport and Social Issues 19, 335–351.

Wann, D.L., 2006. The Causes and Consequences of Sport Team Identification, in: Raney, A.A., Bryant, J. (Eds.), Handbook of Sports and Media. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 331–352.

Wies, S., Moorman, C., 2015. Going Public: How Stock Market Listing Changes Firm Innovation Behavior. Journal of Marketing Research 52, 694–709.

Xie, K., Lee, Y.-J., 2015. Social Media and Brand Purchase: Quantifying the Effects of Exposures to Earned and Owned Social Media Activities in a Two-Stage Decision Making Model. Journal of Management Information Systems 32, 204–238.

Zhang, Y., Pennacchiotti, M., 2013. Predicting purchase behaviors from social media, in: Proceedings of the 22nd International Conference on World Wide Web. WWW '13. Edited by: Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo A. Baeza-Yates, and Sue B. Moon, pp. 1521–1532.

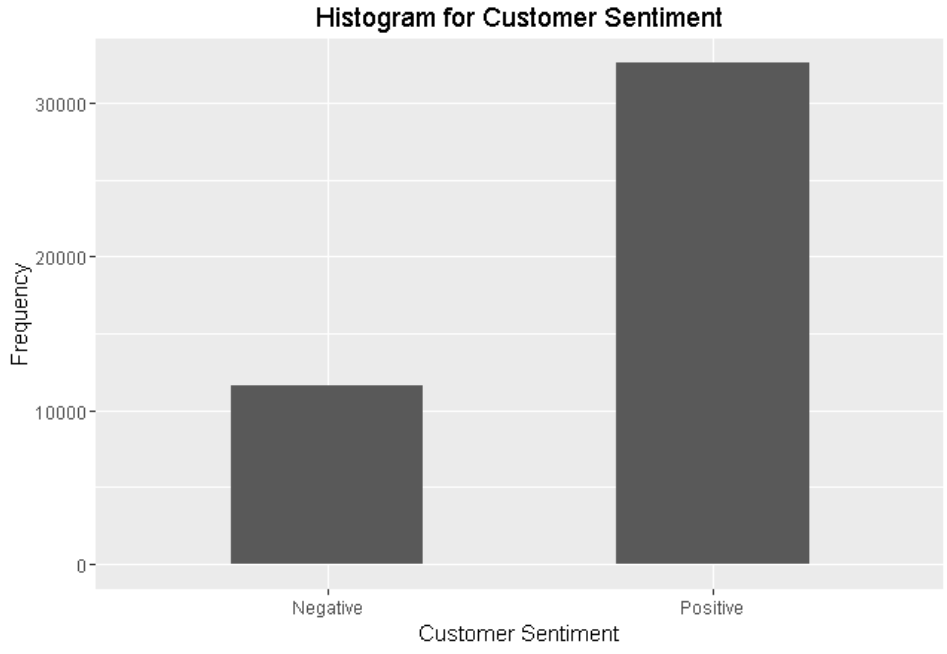
10.Appendices

Appendix A.1: Customer sentiment equation: Descriptive statistics

	MEAN	SD	RANGE
Red Cards	0.15	0.38	[0,2]
Yellow Cards	1.87	1.19	[0,6]
MGC	6.83	7.82	[1,52]
Home match	0.52	0.50	[0,1]
*Likes MGC posts**	0.14	0.22	[0,1.59]
Event Facebook	0.03	0.16	[0,1]
Event attending	0.03	0.17	[0,1]
Customer sentiment$_{-1}$	0.27	0.45	[0,1]
Other UGC valence	0.18	0.55	[-3.6,12]
Other UGC volume	111.73	138.94	[1,1108]
*Comment length**	4.04	1.06	[1.09,8.74]
Result	19% draw, 30% loss, 51% win		
Type match	9% Cup match, 19% European match, 51% normal, 21 % Play-off		
Quality opponent	49% low, 29% medium and 22% high		

*logarithm

Appendix A.2: Customer Sentiment distribution



Appendix A.3: Customer sentiment equation: Correlation matrix of the relevant variables

Red Cards	1.000							
Yellow Cards	0.279	1.000						
MGC	-0.038	0.011	1.000					
Likes MGC posts	-0.069	-0.052	0.264	1.000				
Customer sentiment$_{-1}$	-0.008	-0.004	0.019	0.062	1.000			
Other UGC valence	-0.031	-0.020	0.046	0.084	0.019	1.000		
Other UGC volume	-0.025	-0.050	-0.100	-0.112	-0.029	-0.007	1.000	
Comment length	0.053	0.031	-0.093	-0.225	0.055	-0.054	0.149	1.000

Correlations above absolute value of 0.009 are significant at $p < 0.05$, two tailed

Appendix B: Engagement models estimation

The engagement model estimation is based on the Limdep implementation of the RE sample selection model (Greene, 2016a). Starting with the buying and contribution margin from the 'Method' section:

$$BUY^*_{i,t} = \alpha_{1i} + \beta_1 x_{1i,t} + u_{1i,t}\,, \qquad (B.1)$$

$$BUY^*_{i,t} > 0 \; if \; BUY_{i,t} = 1 \qquad (B.2)$$
$$BUY^*_{i,t} \leq 0 \; if \; BUY_{i,t} = 0\,.$$

$$CM_{i,t} = \alpha_{2i} + \beta_2 x_{2i,t} + u_{2i,t}\,, \qquad (B.3)$$

where α_{1i} and α_{2i} are vectors of customer specific intercepts, β_1 and β_1 are vectors of coefficients, $x_{1i,t}$ and $x_{2i,t}$ are vectors containing the predictor variables and $u_{1i,t}$ and $u_{2i,t}$ capture the error terms in the buying and contribution margin respectively ($u_{1i,t}$ ~ N[0,1] and $u_{2i,t}$ ~ N[0,σ^2]. Let ρ = cor($u_{1i,t}$,$u_{2i,t}$), then the contribution of group *i* to the log likelihood can be described as:

$$\log L_i |\, \alpha_{2i}, \alpha_{1i} = \sum_{BUY_{i,t}=0} \log \Phi(-\alpha_{1i} - \beta_1 x_{1i,t}) + \sum_{BUY_{i,t}=1} \left[\frac{-\log 2\pi}{\pi} - log\,(\sigma) - \frac{\left(CM_{i,t} - \alpha_{2i} - \beta_2 x_{2i,t}\right)^2}{2} + \log \Phi \left[\frac{\left(\alpha_{1i} + \beta_1 x_{1i,t}\right) + \left({}^{\rho}/_{\sigma}\right)\left(CM_{i,t} - \alpha_{2i} - \beta_2 x_{2i,t}\right)}{\sqrt{1-\rho^2}} \right] \right] \qquad (B.4)$$

However, α_{2i} and α_{1i} are unobserved. We therefore obtain the unconditional log likelihood by integrating out the random effects:

$$\text{Let } L_i|\alpha_{1i}, \alpha_{2i} = \psi_{it} \qquad (B.5)$$

$$\text{Then, } \iint g(\alpha_{1i}, \alpha_{2i})\, \psi_{it}\, d\,\alpha_{2i} d\alpha_{1i} \qquad (B.6)$$

In order to solve this, Monte Carlo simulation is used and the integral is approximated by

$$E_{\alpha_{1i},\alpha_{2i}}[L_i|\alpha_{1i}, \alpha_{2i}] \approx \frac{1}{R} \sum_{r=1}^{R} L_i|\alpha_{1ir}, \alpha_{2ir}, \qquad (B.7)$$

where α_{1ir}, α_{2ir} are R random draws from the joint distribution of α_{1ir} and α_{2ir}. The approximation improves with increasing R. The simulation allows for two parameters to be set: the method of random draws and the number of draws. Based on the recommendations in Greene (2016b) we use 1000 Halton draws (for a more in depth discussion of Halton draws, see Greene (2016b) and Train (1999)).

Then, the total log likelihood can be described as:

$$\log L = \sum_{i=1}^{N} \log L_i \qquad (B.8)$$

This likelihood function is then maximized by solving the likelihood equations:

$$\frac{\partial \log L}{\partial \Theta} = \sum_{i=1}^{N} \frac{\partial log L_i}{\partial \Theta} = 0, \qquad (B.9)$$

where Θ refers to the vector of parameters in the model. These derivatives must be approximated as well. Please see Greene (2016b) for a detailed description of the process.

References

Greene, W.H., (2016a), Sample Selection Models for Panel Data, in: Econometric Modeling Guide Limdep 11. Econometric Software, Inc.
-------- (2016b). Random Parameter Models, in: Limdep 11 Reference Guide. Econometric Software, Inc.
Train, K. (1999) 'Halton Sequences for Mixed Logit,' Manuscript, Department of Economics, University of California, Berkeley.

Appendix C: Engagement models without random effects

	Purchase incidence		Contribution margin	
Variables	Estimate	z-score	Estimate	z-score
Intercept	-0.42 ***	-9.82	159.86 ***	53.00
$\widehat{Customer\ Sentiment}$	0.04 ***	4.15	1.61 *	1.85
Share of Engagement	0.00	-0.27	-0.53	-0.58
$\widehat{Customer\ Sentiment}$ Share of Engagement*	0.00	0.23	0.44	0.69
Page Like	0.01	0.62	1.32	0.71
$Purchase_{t-1}$	1.24 ***	37.14		
$PurchaseAmount_{t-1}$			89.07 ***	276.19
Price paid	0.03 *	1.87		
Tenure	0.01	0.92	3.41 ***	3.80
Contact Volume	-0.05 ***	-3.57	-2.66 **	-1.98
*Year2014*Contact Volume*	0.08 ***	3.39	4.73 ***	2.62
Click-through Rate	0.06 ***	5.73	3.81 ***	4.18
Consumption	0.17 ***	10.74	-2.89 **	-2.51
Gender	0.05 **	2.24	7.28 ***	3.92
IMR	0.03 ***	3.30	12.10 ***	15.90
σ		102.5***	612.80	
ρ		0.85 ***	252.13	
Year dummies		*Included*		
AIC		198,186.6		
Log-Likelihood		-99,058.32		

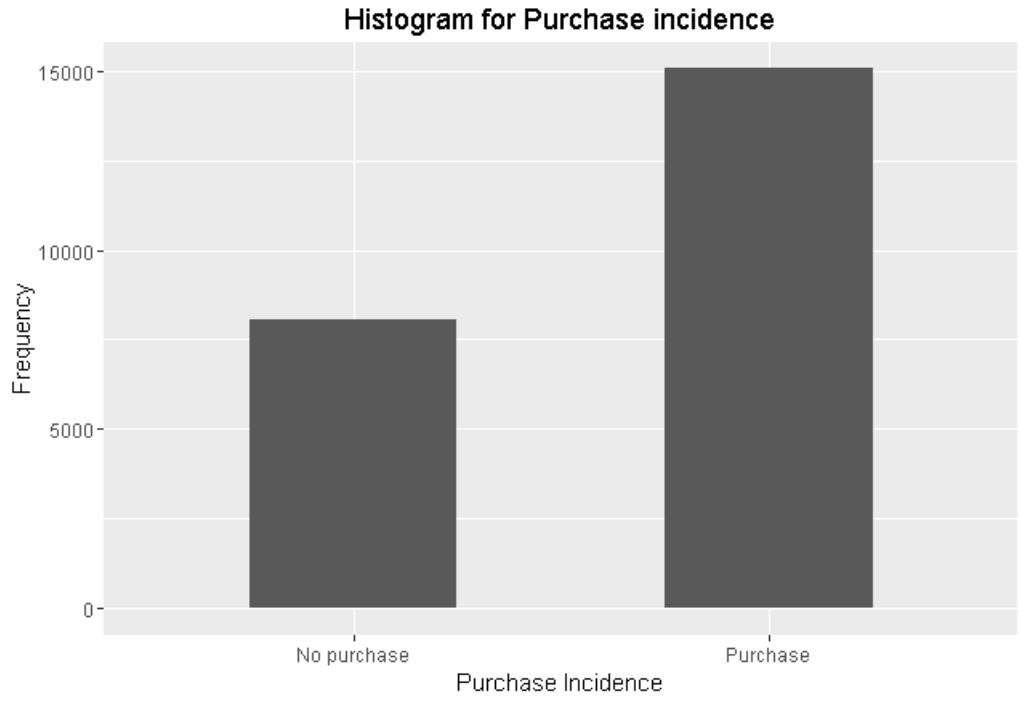
Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized; subscripts are not included for clarity, except for the lagged dependent variables

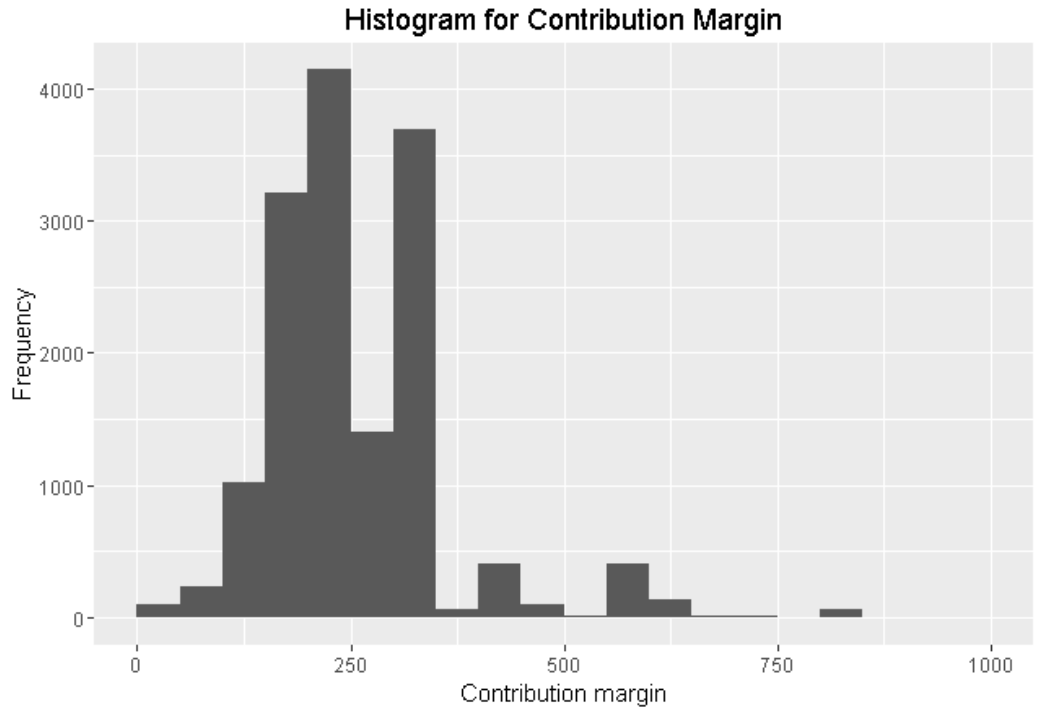
Appendix D.1 : Descriptive summary of the variables in the Engagement model

	PURCHASE INCIDENCE			CONTRIBUTION MARGIN		
VARIABLES	MEAN	SD	RANGE	MEAN	SD	RANGE
$\widehat{Customer\ Sentiment}$	0.64	0.08	[0.30,0.98]	0.64	0.08	[0.30,0.98]
Share of engagement	15.64	20.29	[0,198]	15.67	20.29	[0,198]
Page Like	0.56	0.49	[0,1]	0.58	0.49	[0,1]
Tenure	4.95	3.08	[0,10]	5.22	2.98	[0,10]
$Purchase_{t-1}$	0.64	0.48	[0,1]			
$purchaseAmount_{t-1}$				257.86	125.36	[0,3120]
Price Paid	145.58	130.66	[0,814]			
*Contact Volume**	0.43	0.62	[0,1.65]	0.46	0.64	[0,1.65]
Click-Through Rate	0.02	0.06	[0,1]	0.02	0.07	[0,1]
Consumption	0.26	0.33	[0,1]	0.35	0.35	[0,1]
Gender	0.76	0.42	[0,1]	0.80	0.4	[0,1]

*logarithm

Appendix D.2 : Distribution of purchase incidence and contribution margin





Appendix D.3 : Correlation matrix of the relevant variables in the Engagement model

Correlation for purchase incidence equation

$\widehat{CustomerSentiment}$	1.000						
Share of engagement	0.080	1.000					
Tenure	0.010	-0.008	1.000				
PricePaid	0.033	-0.024	0.270	1.000			
ContactVolume	-0.018	0.057	0.187	0.151	1.000		
Click-through rate	-0.051	-0.018	0.076	0.372	0.320	1.000	
Consumption	-0.077	0.119	0.243	0.384	0.358	0.206	1.000

Correlations above absolute value of 0.013 are significant at $p < 0.05$, two tailed

Correlation for contribution margin equation

$\widehat{CustomerSentiment}$	1.000						
Share of engagement	0.086	1.000					
Tenure	-0.001	-0.026	1.000				
$PurchaseAmount_{t-1}$	0.049	-0.043	0.290	1.000			
Contact Volume	-0.010	0.055	0.166	0.152	1.000		
Click-through rate	-0.057	0.018	0.069	0.394	0.339	1.000	
Consumption	-0.089	0.146	0.271	0.374	0.429	0.214	1.000

Correlations above absolute value of 0.016 are significant at $p < 0.05$, two tailed

Appendix E: Customer Sentiment Equation Results with 1 and 3 day timeframe

	1-day timeframe		3-day timeframe	
Variables	Estimate	z-score	Estimate	z-score
Intercept	0.979 ***	6.626	0.607 **	3.765
Result (Lost)$_m$	-0.349 ***	-4.057	-0.357 ***	-3.907
Result (Won)$_m$	0.894 ***	11.589	1.134 ***	13.664
RedCards$_m$	-0.052 *	-1.747	-0.051 *	-1.647
YellowCards$_m$	-0.041	-1.487	0.004	0.120
TypeMatch (Eur)$_m$	0.085	0.665	0.324 **	2.371
TypeMatch (Nor)$_m$	0.073	0.628	0.091	0.735
TypeMatch (PO)$_m$	0.138	1.083	0.166	1.223
QualityOpponent (Medium)$_m$	-0.001	-0.011	0.122	1.466
QualityOpponent (High)$_m$	0.031	0.416	0.045	0.565
MGC $_{u,c,m}$	0.211 ***	6.052	0.262 ***	4.607
*ResultLost$_m$ * MGC$_{u,c,m}$*	0.427 ***	8.425	0.044	0.595
*ResultWon$_m$ * MGC $_{u,c,m}$*	-0.303 ***	-7.723	-0.263 ***	-4.318
Home Match$_m$	-0.119 **	-2.121	-0.080	-1.325
LikesMGCPosts$_{u,c,m}$	0.254 ***	17.946	0.235 ***	12.956
EventFacebook$_{u,m}$	0.178 **	2.462	0.128	1.367
EventAttending $_{u,m}$	-0.101	-1.510	-0.050	-0.579
Customer Sentiment $_{u,c,m-1}$	0.118 ***	4.840	0.056 *	1.813
Other UGC Valence $_{u,c,m}$	0.122 ***	10.835	0.072 ***	5.243
Other UGC Volume $_{u,c,m}$	0.095 ***	7.114	0.128 ***	7.322
Comment length $_{u,c,m}$	-0.142 ***	-12.228	-0.018	-1.285

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized

Appendix F: Engagement model with weighted predicted Customer Sentiment

	Purchase incidence		Contribution margin	
Variables	Estimate	z-score	Estimate	z-score
Intercept	-0.28 ***	-6.73	165.71 ***	65.11
Customer Sentiment	0.05 ***	4.49	1.70 **	2.04
Share of Engagement	0.00	-0.35	-0.58	-0.69
Customer Sentiment Share of Engagement*	0.00	0.37	0.48	0.79
Page Like	0.02	0.85	1.81	1.12
$Purchase_{t-1}$	1.07 ***	31.61		
$PurchaseAmount_{t-1}$			75.66 ***	303.24
Price paid	0.09 ***	5.46		
Tenure	0.01	1.21	5.63 ***	6.65
Contact Volume	-0.05 ***	-3.53	-2.72 **	-2.21
*Year2014*Contact Volume*	0.07 ***	3.21	3.72 **	2.21
Click-through Rate	0.06 ***	5.81	3.68 ***	4.50
Consumption	0.19 ***	12.54	2.57 **	2.41
Gender	0.06 **	2.46	8.94 ***	5.32
IMR	0.02 **	2.42	14.58 ***	22.85
σ		0.01	0.06	
ρ		0.89 ***	317.23	
Year dummies		*Included*		
AIC		198,046.5		
Log-Likelihood		-98,986.23		

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized; subscripts are not included for clarity, except for the lagged dependent variables

Appendix G: Endogeneity due to social effects

Given that we investigate social effects from observational data, possible identification issues arise due to endogeneity (Manski 1993). The main concern is that social connections show similar behavior not only as a result of tie influence, but due to other reasons, referred to as unobserved correlations (Manski 2000; Nitzan and Libai 2011). This would render our social variables endogenous and could bias our parameter estimates. In this appendix, we show how we attempted to mitigate this issue.

1) The observed social effects can be due to endogenous effect, which refers to the propensity of an individual to behave in some way can vary with the behavior of the group. This is also called the 'simultaneity' or 'reflection' problem in literature (Manski 1993). In our specific case, this refers to the problem that for instance a defecting neighbor or spending affects the defection or spending of the focal customer, and at the same time the focal customer's defection or spending influences the neighbor. In accordance with previous literature (e.g., Iyengar, Van den Bulte, and Lee 2015; Iyengar, Van den Bulte, and Valente 2010; Manchanda, Xie, and Youn 2008; Risselada, Verhoef, and Bijmolt 2013), we mainly use temporal precedence in order to avoid simultaneity, i.e. we model social contagion in terms of lagged rather the contemporaneous peer effects (e.g., our variable contains the number of defectors in period t in order to predict defection in period t +1)[1]. Moreover, the reflection problem is not very likely since we also control for lagged behavior of the focal customer (Iyengar, Van den Bulte, and Lee 2015).
2) There may be confounding environmental factors (contextual effects), in which the propensity of an individual to behave in some way is influenced by unobserved factors (external shocks) that also influence the group varies with the exogenous characteristics of the group or external shocks to which the group is exposed (Aral, Muchnik, and Sundararajan 2009). Examples in our specific case can be the absence of the championship title or good player transfers. Time specific variables, included in our model, can capture these external shocks.
3) Third, there may be correlated or social effects, which reflects the tendency of customers in a group to behave similarly because of similar individual characteristics. In order to account for these effects, previous models have incorporated variables that indicate similarity such as demographics (Nair, Manchanda, and Bhatia 2010; Nitzan and Libai 2011), or by including a random effect specification for heterogeneity (Hartmann et al. 2008). Thus, in our model we also control for these effects by including demographics and random effects per customer.

[1]Note that the implicit assumption is made that the customers are not forward looking with regard to their own and other's behavior

References

Aral, Sinan, Lev Muchnik, and Arun Sundararajan (2009), "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, 106 (51), 21544–49.

Hartmann, Wesley R., Puneet Manchanda, Harikesh Nair, Matthew Bothner, Peter Dodds, David Godes, Kartik Hosanagar, and Catherine Tucker (2008), "Modeling social interactions: Identification, empirical methods and policy implications," *Marketing Letters*, 19 (3–4), 287–304.

Iyengar, Raghuram, Christophe Van den Bulte, and Jae Young Lee (2015), "Social Contagion in New Product Trial and Repeat," *Marketing Science*, 34 (3), 408–29.

———, ———, and Thomas W. Valente (2010), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30 (2), 195–212.

Manchanda, Puneet, Ying Xie, and Nara Youn (2008), "The Role of Targeted Communication and Contagion in Product Adoption," *Marketing Science*, 27 (6), 961–76.

Manski, Charles F. (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60 (3), 531–42.

——— (2000), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, 14 (3), 115–36.

Nair, Harikesh S, Puneet Manchanda, and Tulikaa Bhatia (2010), "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research*, 47 (5), 883–95.

Nitzan, Irit and Barak Libai (2011), "Social Effects on Customer Retention," *Journal of Marketing*, 75 (6), 24–38.

Risselada, Hans, Peter C. Verhoef, and Tammo H.A. Bijmolt (2013), "Dynamic Effects of Social Influence and Direct Marketing on the Adoption of High-Technology Products," *Journal of Marketing*, 78 (2), 52–68.

Appendix H: Engagement model with network variables

Table H1: Purchase incidence equation

	Model 1: only SM related variables		Model 2: SM and network info		Model 3: complete model	
Variables	**Estimate**	**z-score**	**Estimate**	**z-score**	**Estimate**	**z-score**
Intercept	0.63 ***	21.63	0.68 ***	23.60	0.28 ***	3.80
Customer Sentiment	0.08 ***	3.27	0.10 ***	5.26	0.09 ***	4.95
Share of Engagement	-0.01	-0.43	-0.02	-0.95	-0.01	-0.75
*SoE * Customer Sentiment*	0.00	0.06	-0.00	-0.04	-0.00	-0.12
Page Like	0.12 ***	3.18	0.14 ***	4.30	-0.01	-0.16
Network Customers			0.38 ***	17.33	0.17 ***	8.94
Network Defection			-0.01	-0.69	-0.04 ***	-2.59
Homophily			0.01	0.38	0.01	0.58
Purchase$_{t-1}$					0.98 ***	17.01
Price Paid					0.05 **	2.12
Tenure					0.00	0.16
Contact Volume					-0.06 **	-2.57
*Year2014*Contact Volume*					0.03	0.73
Click-through Rate					0.10 ***	5.35
Consumption					0.36 ***	14.89
Gender					-0.01	-0.14
IMR	0.080 ***	8.66	0.100 ***	6.53	0.01	.93
Year dummies	*Included*		*Included*		*Included*	

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized; subscripts are not included for clarity, except for the lagged dependent variables

Table H2: Contribution margin equation

	Model 1: only SM related variables		Model 2: SM and network info		Model 3: complete model	
Variables	**Estimate**	**z-score**	**Estimate**	**z-score**	**Estimate**	**z-score**
Intercept	182.55 ***	111.97	183.97 ***	117.16	192.52 ***	76.56
$\widehat{Customer\ Sentiment}$	3.90 ***	5.00	3.56 ***	4.75	1.65 **	2.05
Share of Engagement	-0.49	-0.66	-1.65 **	-2.28	-0.95	-1.23
*SoE * $\widehat{Customer\ Sentiment}$*	-0.05	-0.09	-0.11	-0.18	-0.67	-1.07
Page Like	4.63 ***	3.38	3.94 ***	2.95	-0.31	-.020
Network Spend			3.84 ***	4.90	6.20 ***	12.84
$PurchaseAmount_{t-1}$					33.39 ***	61.73
Tenure					6.22 ***	7.89
Contact Volume					-1.44	-1.30
*Year2014*Contact Volume*					0.75	0.44
Click-through Rate					2.58 ***	3.86
Consumption					6.28 ***	5.88
Gender					6.49 ***	3.69
IMR	19.41 ***	38.46	18.31 ***	33.79	11.15 ***	18.08
σ	0.02	0.09	0.022	0.09	0.02	.08
ρ	1.00 ***		1.00 ***	******	0.90 ***	156.8
Year dummies	*Included*		*Included*		*Included*	
AIC	86,951.6		86,832.7		85,082.2	
Log-Likelihood	-43,453.81		-43,390.34		-42,500.10	

Note: * p<0.1, ** p<0.05, *** p<0.01; coefficients are standardized; subscripts are not included for clarity, except for the lagged dependent variables

4

The Added Value of Social Media Data in B2B Customer Acquisition Systems: A Real-life Experiment

4. The Added Value of Social Media Data in B2B Customer Acquisition Systems: A Real-life Experiment

Abstract

Business-to-business organizations and scholars are becoming increasingly aware of the possibilities social media and predictive analytics offer. Despite the interest in social media, only few have analyzed the impact of social media on the sales process. This paper takes a quantitative view to examine the added value of Facebook data in the customer acquisition process. In order to do so, we devise a customer acquisition decision support system to qualify prospects as potential customers, and incorporate commercially purchased prospecting data, website data and Facebook data. Our system is subsequently used by Coca Cola Refreshments Inc. (CCR) to generate calling lists of beverage serving outlets, ranked by their likelihood of becoming a customer. In this paper we report the results, in terms of prospect- to- customer conversion, of a real-life experiment encompassing nearly 9,000 prospects. The results show that Facebook is the most informative data source to qualify prospects, and is complementary with the other data sources in that it further improves predictive performance. We contribute to literature in that we are the first to investigate the effectiveness of social media information in acquiring B2B-customers. Our results imply that Facebook data challenge current best practices in customer acquisition.

This chapter is based on the published article Meire, M. , Ballings, M. and Van den Poel, D. (2017). The Added Value of Social Media Data in B2B Customer Acquisition Systems: A Real-life experiment, Decision Support Systems, 104, December 2017, p26-37.

1. Introduction

While social media have given rise to a vast body of literature in marketing (e.g., Goel and Goldstein, 2013; Goh et al., 2013; Kumar et al., 2015; Xie and Lee, 2015), most of this research focuses on business-to-consumer (B2C) applications. Within business-to-business (B2B) environments, the potential of social media has already been recognized, but the adoption of social media is slower compared to B2C companies (Michaelidou et al., 2011). Existing literature describes in a qualitative way how social media can be used, mainly within a B2B selling process or relationship. However, any formal model or analysis of the abundance of social media data in a B2B environment is lacking.

The magnitude of these social media data becomes most apparent if we look at some summary figures. Facebook[3] contains over 60 million company pages and 1.79 billion active user profiles interacting with these pages at the end of 2016 (Facebook, 2016; VentureBeat, 2016), and serves as a prime example of big data (Wedel and Kannan, 2016). These magnitudes of new (e.g., voice, text, photo and video) data bring along new challenges. Indeed, the Marketing Science Institute (MSI) lists as one of its research priorities for 2016-2018 "New data, new methods, and new skills- how to bring it all together?" with key issues described as: "How to bring multiple sources and types of information together […] to make better decisions […].", "Integrating big data analysis with managerial decision making." and "New approaches and sources of data – what are the roles of artificial intelligence, […], machine learning?" (MSI Research Priorities 2016-2018, 2016). According to Lilien (2016), there is also a spiking interest of B2B selling firms for machine learning and predictive analytics, driven by new data sources that become available. In summary, several authors have stated the need to explore the added value of big data applications and analytics in business environments, thereby taking into account the data, tools and algorithms that can be used (e.g., Baesens et al., 2016; Wedel and Kannan, 2016). Recently, Chen et al. (2015) showed that the use of big data analytics was responsible for 8.5% explained variance in asset productivity and 9.2% explained variance in business growth, which indicates the relevance of big data for value creation.

We add to existing literature concerning B2B social media usage by incorporating social media within a B2B customer acquisition decision support system. In the history of customer relationship management (CRM), the acquisition process has received less attention compared

[3] We chose Facebook as our focus of analysis as this is by far the largest network in terms of users and available variables and is named as one of the 'big three' in 'big data' (Leverage New Age Media, 2015)

to retention and customer lifetime value (CLV). The underlying reasons are that the customer acquisition process is more complex, less data of poorer quality are available, and customer acquisition is typically more expensive compared to retention campaigns (Reinartz and Kumar, 2003). The rise of social media can be conceived of as an opportunity to obtain a better defined profile of prospects, thereby allowing to create better customer acquisition prediction models. Specifically, we evaluate the predictive value of data extracted from the prospects' social media page (Facebook pages), and compare it with data extracted from their website, and data that the focal company buys from a specialized vendor. We implement this research using a real-life experiment with Coca Cola Refreshments USA Inc. (CCR) in which we had CCR's call center call nearly 9,000 prospects. Prospects in this particular case refer to on premise beverage-serving companies such as bars and restaurants, which we call outlets from hereon.

The main contributions of this paper are: 1) We posit, evaluate and assess a customer acquisition decision support system on a large scale and show the financial benefits of this new approach using a real- life experiment with Coca Cola Refreshments USA, 2) We add to the existing B2B social media literature by taking a quantitative, big data view on social media instead of a qualitative one and 3) We add to the existing B2B acquisition literature by incorporating a new, freely available data source over established data sources for better prediction models.

In the next section, we will first review the B2B acquisition process, previous literature on social media in a B2B environment and the potential added value of social media for B2B customer analytics. Next, we describe our data sources, along with the methodology. This methodology is evaluated in a real-life experiment in the Results section. Subsequently, we provide a discussion of the results and the implications for business implementations. The final section addresses limitations and outlines future research.

2. Literature review

2.1. B2B acquisition framework

The customer acquisition process is a very complex process, especially in a B2B environment. Organizations' buying decisions are taken by a group of people, often called the Decision Making Unit (DMU), and rely on budget and cost considerations (Webster and Wind, 1972). Typically, the process is split up in different stages. We follow the approach outlined in D'Haen and Van den Poel (2013). Their 'sales funnel' consists of four stages. In the first stage, there is only a list of suspects. These are all potential new customers (D'Haen and Van den Poel, 2013).

In most industries, a complete list of potential customers does not exist and in this case the list should be thought of as an ideal. Subsequently, this initial list is reduced to a list of prospects that can be identified. This is the stage where most companies start the sales process, either with an acquired list from a specialized vendor (Blattberg et al., 2008) or with a list obtained from the marketing department (Sabnis et al., 2013). The third stage consists in qualifying these prospects, which yields a list of leads. Typically, in practice, qualifying prospects is based on intuition, gut feeling and simple rules (Jolson, 1988; Monat, 2011). However, more informed approaches exist as explained in Blattberg et al. (2008): profiling, random testing of prospect lists, a two-step acquisition model and regression models. These approaches have proven their usefulness in several applications (e.g., D'Haen et al., 2016; Reinartz et al., 2005; van Wangenheim and Bayón, 2007). Finally, in the fourth stage of the sales funnel, the lead is converted to a real customer.

Similar to the complexity of the sales process, the modeling of this process can be seen as a complex undertaking. Indeed, D'Haen and Van den Poel (2013) point out the iterative nature of the sales process. In a first phase, there is only information available on customers versus prospects. Hence, a type of profiling method is used, identifying prospects that look similar to existing customers. Each prospect receives a score that reflects the probability to become a customer. Subsequently, this list of prospects is given to the sales team. The second phase starts when feedback on the first list of prospects is received (D'Haen and Van den Poel, 2013). This feedback can take various forms, depending on the stage of the acquisition process that the company is interested in. Examples are the qualification of the prospects as good or bad leads, prospects entering a sales conversation or not, and the closure of a deal or not. Which definition of feedback is most suitable depends on the nature of the business, the time window and the resources of the company: information on the closure of a deal is the most interesting type of feedback to a company, but given the long sales cycle in B2B-sales (Kumar et al., 2013), it may be more effective to use the qualification as good or bad leads as feedback. This feedback gives the opportunity to model the second phase in which the 'good' prospects are modeled versus the 'bad' prospects, in terms of the feedback received. Finally, this process is iterative as the model can be re-estimated and refined each time new feedback becomes available (D'Haen and Van den Poel, 2013). In this paper we apply this iterative model on a large-scale real-life case study, thereby helping to validate this model. In Phase I, we estimate and evaluate the quality of the probability of prospects to become a customer, based on the similarity with customers.

In Phase II, with feedback data available, we model which prospects will be converted into customers, based on information from previous successful conversions (Reinartz et al., 2005).

2.2. Social media in a B2B sales context

Several authors have tried to obtain more insight into the reasons of success of an acquisition attempt (e.g., Walker et al., 1977; Weitz et al., 1986; Zoltners et al., 2008), and most of this research focuses on the antecedents of salespersons' performance. Weitz et al. (1986) mention the capabilities of a salesperson, driven by knowledge and information acquisition skills, as important factors. More recent work stresses the adoption of information technology by the sales force (Ahearne et al., 2008; Schillewaert et al., 2005), and shows a positive relationship between the use of IT and sales performance mediated by the positive influence of IT on knowledge and adaptability of the salesperson. Moreover, Zoltners et al. (2008) show that data and tools available to the sales team are one of the drivers of sales force effectiveness and are seen as one of the high impact opportunities for sales teams by both practitioners and academics. With the recent rise of social media as a new data source, the use of social media within a B2B context thus provides new opportunities to improve sales force effectiveness. The (B2B) sales process becomes more and more influenced by the internet and more specifically, social media (Marshall et al., 2012). While Michaelidou et al. (2011) mention that that the adoption of social media by B2B companies is slower compared to the B2C markets, the usefulness of social media in a B2B context has already been recognized by several scholars. Giamanco and Gregoire (2012) suggest three stages in which social media can be used. These stages are prospecting (i.e., finding new leads), qualifying leads, and managing relationships. In the first stage, sales representatives use social media to identify potential buyers. In the second stage, the quality of these leads is examined using information available on social media (e.g., 'Does this person have the authority to buy?', 'Do they have a budget?' (2012). Finally, social media can be used to manage the relationships with existing customers. The social media they refer to are LinkedIn, Twitter and Facebook. Similarly, Rodriguez et al. (2012) identified a three step process using social media: creating opportunity, understanding customers and relationship management. It is clear that these steps are linked to the previous ones and the main difference is that the relationship stages are expanded over several categories. Creating opportunity embraces both the prospecting and qualifying stages of Giamanco and Gregoire (2012). Moreover, these authors show that social media usage has a positive effect on the results of prospecting and qualifying activities (Rodriguez et al., 2012). Finally, Andzulis et al. (2012) state that social media can and should be integrated into the entire sales process.

These papers share the common idea that social media are important in a B2B selling context. They posit ideas and frameworks and elaborate on how salespeople can identify new prospects, on how they can use social media to identify the good prospects and how social media can be used to start or maintain the relationship with the customer. Social media are recognized as a tool to make the sales process less costly and more effective and are seen as an extension of traditional customer relationship management (CRM), leading to Social CRM activities (Rodriguez et al., 2012; Trainor et al., 2014). By building rapport with the prospective customer, the accuracy of the sales process is expected to increase.

While the papers mentioned in the previous paragraph have in common that they highlight the importance of social media, they also share some limitations. Most of the papers focus on identifying and qualifying procurement officers of prospective companies. This is a generalizing view on the sales process, which may not always be suitable. First, while the focus on individual members of a DMU is necessary for complex products and buying organizations, Homburg et al. (2011) indicate that the customer orientation is dependent on the standardization of the product, the importance of the product and competitive intensity. Thus, this suggests that such a degree of customer knowledge is not required for certain products or markets (Verbeke et al., 2008) and would even lower overall sales performance in these cases (Homburg et al., 2011). Second, in many cases the prospects or leads are delivered by the marketing department (Sabnis et al., 2013) based on lists from specialized vendors, which reduces the need to identify prospects based on social media. Moreover, the process of identifying and qualifying leads is a very time consuming process, in terms of searching and evaluating the available information. Sabnis et al. (2013) mention that there are already a lot of competing demands on the sales representative's time. Verifying social media profiles of generated leads would thus not be probable either, and the literature does not mention whether or where social media can otherwise help to solve this issue. All in all, we feel that the current qualitative focus on social media in the literature ignores important opportunities, related to the big data nature of social media.

With this research, we aim to overcome these limitations and take a different view on the use of social media in the sales process by looking at social media as 'big data' (Baesens et al., 2016). We will focus specifically on the 'qualifying' stage of the sales process. First, we focus on company characteristics instead of specific buyer information by using companies' social media pages. This approach is justified by the standardization of the product of the B2B company studied, Coca-Cola Refreshments USA (Homburg et al., 2011), and the fact that we

are dealing with bars and restaurants in which the DMU is mostly restricted to one person (the owner). Second, we use an automatic approach to collect and process information, eliminating the manual screening of social media profiles and thus freeing up time for other activities. Third, we determine the usefulness of social media to reduce the prospect list to a greatly reduced list of leads, which are worth pursuing by sales representatives. In sum, we move social media use in a B2B context from a purely describing, qualitative view to a data-oriented which uses information systems to collect, clean and analyze the data based on machine learning techniques.

2.3. Social media as a data source

The main challenge when qualifying leads is the lack of qualifying characteristics (Järvinen and Taiminen, 2016). Indeed, for prospect scoring, the seller can only rely on data that is either publicly available or available for purchase, as there is no formal relationship with the prospect yet. This data is, however, not always relevant or informative with respect to the prospect's interest in the product (Long et al., 2007). Therefore, from a big data perspective, it is important to gather different data sources and apply algorithms to filter out relevant information. Firms have started to realize this and are now collecting huge amounts of data from diverse sources to increase prediction model performance (Lilien, 2016). We collect data from three sources: commercially purchased data, websites and social media. We hypothesize social media to be the richest source of information when compared to websites and commercially purchased data, based on three advantages of social media.

First, commercial data from specialized vendors is a very expensive source of information, given that these lists tend to be of poor quality (D'Haen et al., 2013) as they often provide 'best estimates' of data (e.g., estimates of revenue (Laiderman, 2005)) and contain a lot of missing values (D'Haen et al., 2016). Websites and social media pages, however, are generated by the company mainly to provide information to customers or other stakeholders (D'Haen et al., 2016). In that respect, companies benefit from providing correct and complete information, making this a more reliable source of information (Melville et al., 2008). Previous research had already shown that website data provide better estimations compared to commercially available data (D'Haen et al., 2013).

Second, we reason that social media pages also have advantages over websites, since the information on social media pages is updated more frequently (e.g., regular posts on a Facebook

page) and the information comes in a standardized format (e.g., JSON files extracted using the API) versus the unstructured text on websites, which makes it more difficult to analyze.

Finally, we believe that different information types are available on social media. Yu and Cai (2007) indicate three types of data that help qualifying B2B customers: company characteristics, customer behavior and attitudinal information. The customer's company characteristics indicate the business background, the size of the company, the geographic location and product range, amongst others. Customer behavior includes transaction records of the customer with the company. Attitudinal information includes the attitudes of the customer-company towards its vendors, personnel, service and customers, and the vision of the customer-company. Customer behavior is not available for prospects and can be ignored for this analysis. Commercial data typically contain company characteristics (Laiderman, 2005), and D'Haen et al. (2016) mention that websites provide similar information compared to commercial data, but more complete. Next to company characteristics, we argue that Facebook pages do contain attitudinal information such as the attitude and communication of a prospect towards and with its own customers, the vision of the prospect and popularity (in terms of the number of likes or visitors), and reviews about the prospect. Indeed, the corporate brand can be build and sustained using Facebook pages (Brito Pereira Zamith et al., 2015). It can be argued that this extra information provides more detailed insights into the prospect organization, as similar company 'personalities' (Keller and Richey, 2006) can provide an extra dimension of knowledge over company characteristics. Given the rich information present in social media data, we hypothesize social media data to be most predictive for customer acquisition, as data quality is the best driver to boost predictive model performance (Baesens et al., 2016).

As a conclusion, we summarized the relevant literature concerning customer acquisition in Table 4.1. This table helps to highlight the three main contributions of this paper to extant literature as outlined in the introduction.

Table 4.1: Literature review on Customer Acquisition

Study	Focus	Research Type	Data type			Objective	Industry	Key insight
			Comm.	Web	SM			
Michaelidou et al. (2011)	B2B	Exploratory			X	The use of social network systems for branding	Multiple	B2B companies mainly use SNS as a way to attract new customers
Giamanco and Gregoire (2012)	B2B	Descriptive			X	Review the role of social media in the sales process	/	Show how social media can add value in each step of the sales cycle
Rodriguez et al. (2012)	B2B	Exploratory			X	Influence of social media on selling activities	25 industries	Social media has a positive relationship with sales processes and relationship sales performance
Andzulis et al. (2012)	B2B	Descriptive			X	Review the role of social media in the sales process	/	Show how social media can add value in each step of the sales cycle
Marshall et al. (2012)	B2B	Exploratory			X	Influence of social media on selling activities	26 industries	Contemporary selling is driven in large measure by social media
Trainor et al. (2014)	B2B	Exploratory			X	Influence of social media on selling activities	Multiple	Social media usage positively relates to customer relationship performance
Lix et al. (1995)	B2C	Predictive	X			Prediction of prospects with high propensity to buy	Multiple	Prospects with only commercial data can be scored accurately
Hansotia and Wang (1997)	B2C	Predictive	X			Financially determine which prospects should be contacted	/	Decision to contact a prospect also depends on the contact strategy and CLV
Reinartz et al. (2005)	B2B	Predictive	X			Balancing resources between customer acquisition and retention	High-tech	The decision how much to spend for each customer on acquisition and retention are interrelated
Wangenheim and Bayon (2007)	B2B and B2C	Predictive				Better allocate marketing resources based on the link WOM-customer acquisition	Energy retailer	Word-of-mouth influences new customer acquisitions
Thorleuchter et al. (2012)	B2B	Predictive		X		Integration of web data for customer acquisition	Mail-order	Website data help to make profitability predictions for new customers
D'Haen et al. (2013)	B2B	Predictive	X	X		Evaluation of the best data source	Mail order	Website data are more predictive compared to commercial data
Goel and Goldstein (2013)	B2C	Predictive			X	Show predictive value of social data	Recreational league	Social data add over demographics-only models
D'Haen et al. (2016)	B2B	Predictive	X	X		Evaluation of knowledge data in a multilingual setting	Energy retailer	Expert knowledge adds significantly to the prediction abilities
Our study	B2B	Predictive	X	X	X	Evaluation of the added value of social media data	Beverage consumption	Social media data add value over website and commercial data

3. Methodology

3.1. Data

Our literature review indicates that external data sources are crucial to obtaining information on prospects. Indeed, the company does not have rich transactional data (e.g. sales data) available from prospects as they have not been customers yet. We employ three types of data sources: purchased commercial data from a specialized vendor, data from the prospects' web pages and data from the prospects' Facebook pages. Given the importance of these data sources, we will discuss each of them in more detail below. Data collection started with the commercial data, as this was available for all prospects and customers, and we took a random subsample of 92,900 instances. Next, we looked for the websites of these companies, which resulted in 65,391 records with available websites. Finally, we identified the Facebook pages and end up with 26,622 companies for which all data where available. This data set consisted of 17,536 existing customers and 9,086 prospects. These were used as input for Phase I. Phase II only uses the prospects and thus has a total input of 9,086 observations. We summarize all variables in Appendix A. For the categorical variables, we include (a range of) proportions in Appendix A, while we provide descriptive statistics of the continuous variables in Appendix B.

3.1.1. Commercial Data

The commercial data were acquired from a specialized vendor by the focal company, CCR. However, this list of companies mainly served to identify prospects, ignoring the available information to score the prospects based on a formal model. The type of information included in the commercial data are company size (sales volume, number of employees, square footage and number of PCs), industry type (NAICS-code and further industry sub classification) and other business demographics (women owned, ethnic background of owner, spoken language of owner, homebased business, credit score, franchise indicator, region and related census data for the region). The website of the prospects was also available. In total, 67 variables were created from the commercial dataset, all dummy variables.

3.1.2. Web Data

As a second source of information, we use the publicly available websites of the prospect companies. Therefore, we developed software to crawl the website information of all prospects (the website addresses were present in the commercial dataset). Subsequently, the unstructured information is turned into usable features by applying text mining techniques to the website text. We follow the standard procedures in text mining (Meire et al., 2016). First, raw text

cleaning is applied in combination with stop word removal. Second, a document-term matrix is produced. This matrix links a website to all the words that occur on the website, which results in a sparse matrix not useful for modeling purposes. Hence, we apply Latent Semantic Indexing (LSI) (Deerwester et al., 1990), a technique that allows to reduce the dimensionality of the feature space. This technique uses Singular Value Decomposition (SVD) to reduce the document-term matrix to its first *k* singular vector directions. Given that most of the variance is captured within the first singular vectors, this method reduces the need to include many predictors while keeping most of the variance. We use the first 50 singular vectors in all subsequent analyses.

Based on recommendations of D'Haen et al. (2016), we also include expert knowledge, which is defined as the information that is deemed important by the salespersons based on previous experience. This expert knowledge consists of links to the contact form and social media (Facebook, Twitter and Instagram) on the website. Indeed, one drawback of the LSI method is that specific information may be lost, which is solved by incorporating these specific features directly in the models. Thus, in total, we use 53 (50 singular vectors + 3 expert knowledge) features from the website text.

3.1.3. Facebook Pages

The third source of information consists in Facebook pages. This refers to all information about a prospect that can be found on the Facebook page of a prospect. This Facebook page is a publicly available web page within the Facebook social network, set up by the prospect, in order to communicate to and connect with clients.

In a first stage, we need to identify the Facebook pages of the companies. We set up an information system consisting of two steps. First, we set up a smart searching algorithm that searches for the prospect's Facebook page using the name of the company, the address and the website address. In a second stage, we extracted information from the Facebook pages using the publicly available Facebook API and software that we custom developed. The information comes as a JSON-file, making processing easy and fast compared to the processing of unstructured textual information from websites.

The data drawn from the Facebook page can be divided into the two broad categories that we outlined in the literature review, company characteristics and attitudinal information. First, the Facebook page contains company characteristics such as the price range, industry category and services. Furthermore, we include dummy variables indicating how complete the Facebook

page is (e.g., phone, webpage and location). Second, the Facebook page contains attitudinal information. We include communication of the company with clients such as the number of posts on the Facebook page, the time between two posts and the number of comments, likes and shares of these posts. Moreover, we add measures such as the number of likes, the number of check-ins and the number of messages in which the company was 'tagged' to include popularity measures of the company. In total, 99 variables were created based on Facebook input.

3.2. Models

3.2.1. Phase I models

In line with the theoretical foundations laid out in the literature review, we build two models. First, we start with an initial model that assumes no knowledge about converted versus non-converted prospects (Phase I). This is called the look-alike model or profiling (Blattberg et al., 2008; Lilien, 2016), because we identify 'good prospects' based on the similarity of their characteristics with existing clients. We specify our dependent variable based on the current status of the outlet (i.e., customer vs prospect). Subsequently, we model the dependent variable as a function of our independent variables of commercial, website and Facebook information and derive the propensity that the prospect belongs to the customer group. We use the Random Forest (Breiman, 2001) algorithm to perform the classification task. Next, we rank the prospects based on their predicted score, which represents their probability of becoming a customer. This approach has several advantages over unsupervised learning methods commonly used for look-alike models. First, the Random Forest algorithm is more appropriate compared to unsupervised learning and other supervised learning algorithms given the high dimensionality of the problem, as it is robust to overfitting (Schwartz et al., 2014). Moreover, Random Forest does not assume a linear relationship between the predictors and the dependent variable, which is a desirable feature when working with textual data. Finally, Random Forest has been shown to be among the best all-round classification techniques (Fernández-Delgado et al., 2014; Lash and Zhao, 2016), next to for instance Support Vector Machines or Artificial Neural Networks.

3.2.2. Experiment

The result of the first stage is a list (or multiple lists based on the different datasets tested), ranking the prospects from high to low probability of becoming a customer. This list is passed back to Coca Cola Refreshment's call center to set up the call action. Importantly, in order to avoid any bias in the results, we provided the call center with a non-ranked list and without

prediction score (D'Haen et al., 2016). In order to control this, we calculated the correlation between the historical performance of the sales persons and the percentage of top-10, top-20 and top-50 ranked prospects assigned to each of the salespeople. The resulting correlations are -0.07, -0.09 and -0.13, respectively, which illustrate that the prospects were indeed randomly assigned. Moreover, all prospects were contacted by telephone and using standardized calling scripts. This is important for comparison of the results, as Hansotia and Wang (1997) show that the offer characteristics may influence response behavior.

CCR agreed to call all prospects on the list, including low-ranked prospects. This has the advantage that we gain insight into the overall model performance and the shape of the lift curve, compared to calling only the top x-percentage of the list. This is interesting from both a practical point of view (e.g., to evaluate what is the optimal number of prospects to call or visit (Verbeke et al., 2012)) as well as for future research and modeling considerations. Moreover, it allows to have more training data available for a (presumably) better Phase II model. After six months, we evaluate whether the called prospects were eventually converted to customers or not. In total CCR called 9086 prospects.

Based on the results of this large-scale experiment, we can measure the performance of our models, i.e. do the higher ranked prospects, as identified by the model, have a higher conversion-to-customer rate compared to lower ranked prospects? The performance measures used are explained in the Section Model Performance.

3.2.3. Phase II models

In addition to the ability to measure performance of the Phase I models, the experiment also triggers Phase II of the customer acquisition framework. Indeed, we now have information available from prospects that were converted versus prospects that were not converted, which allows us to estimate more specific models. We use the same independent variables as in Phase I (purchased, website and Facebook variables), and we will also use a Random Forest model. However, we will now model converted versus non-converted prospects.

In each of the two Phases, we make different models comprising different data sources. We distinguish the following models: model 1 (only commercial data), model 2 (only website data), model 3 (only Facebook data), model 4 (commercial + website data), model 5 (commercial + Facebook data), model 6 (website + Facebook data) and model 6, which comprises all information. Finally, following common practice in predictive modeling (Lash and Zhao, 2016), we report 'out-of-sample' estimates of predictive performance. We use five times

twofold cross-validation (5*2 CV, Dietterich (1998)) in order to sort out the impact of having different training and test sets. Figure 4.1 gives a schematic overview of the different models that were estimated for this analysis.

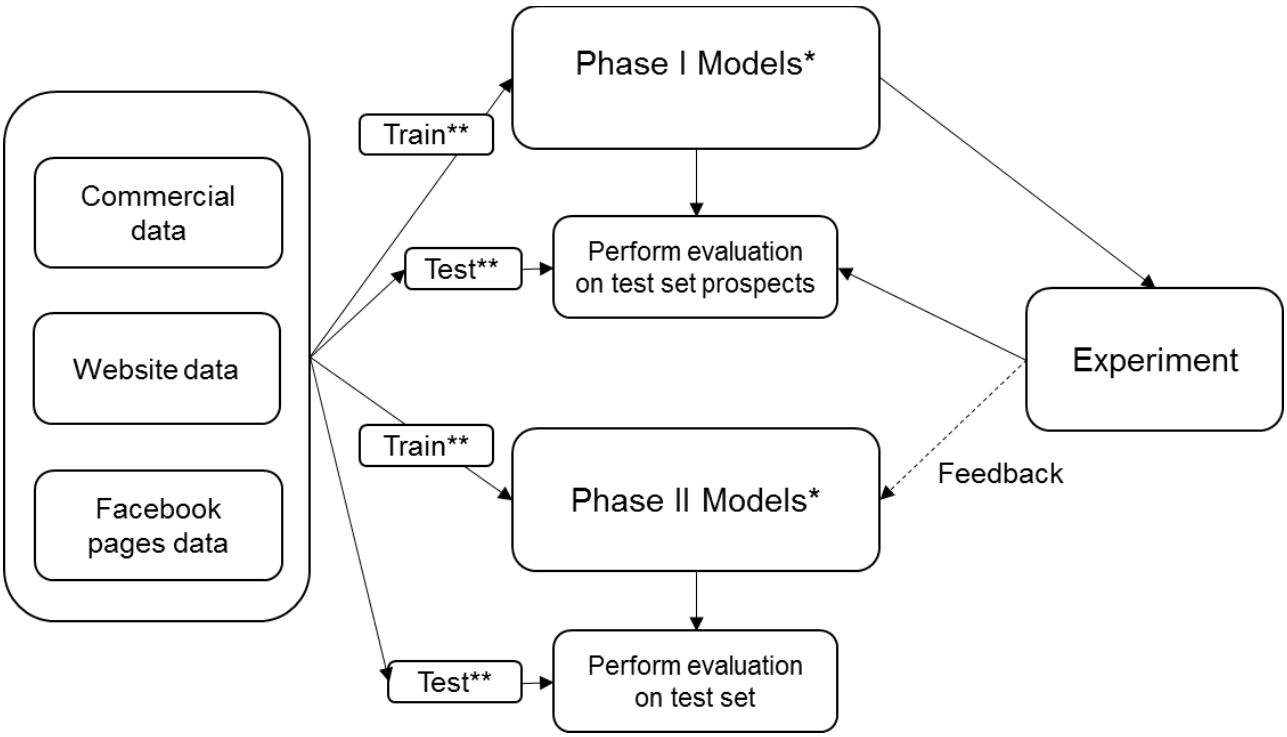


Figure 4.1: Schematic overview of the methodology

** Several models are made (depending on the data sources used), which are not depicted for clarity*
*** The train and test sets of Phase I and Phase II are not the same, as Phase I includes both customers and prospects and Phase II only includes prospects. The train and test sets of Phase II are thus subsets of the Phase I train and test sets, allowing comparison of the results of the prospects in the test set. Moreover, following our cross-validation procedure, we each time have 10 training and test sets.*

3.3. Model Performance

We will evaluate model performance using two widespread measures for classification algorithms, AUC and lift over random selection (Martens et al., 2016). AUC (Area Under the Receiver Operator Curve) is defined as the probability that a randomly chosen positive observation is scored higher than a randomly chosen negative observation. Formally, it can be defined as:

$$AUC = \int_0^1 \frac{TP}{P} d\frac{FP}{N}, \tag{4.1}$$

with *TP* and *FP* true positives and false positives, respectively; *P* the number of observed positive observations and *N* the number of observed negative observations (positive in our models refers to a customer in Phase I and a converted prospect in Phase II). While AUC

measures the performance over the entire range of predictions, lift focuses on the observations with the highest predicted probabilities. Lift over random selection is defined for a certain threshold, which is the top *x*-percentage of the prospects that will be targeted. The top *x*-lift is then defined as the ratio of the percentage of positive cases in the top x-percent scored prospects and the overall percentage of positive cases and is calculated by the following formula:

$$Top\ x - lift = \frac{\frac{P_{top\ x}}{P_{top\ x} + N_{top\ x}}}{\frac{P}{P+N}}, \tag{4.2}$$

where $P_{top\ x}$ and $N_{top\ x}$ are the number of positives and negatives in the top x-percent, respectively and P and N are the number of positives and negatives in the entire sample, respectively. Instead of focusing on top *x*-lift, we will plot the lift curve, plotting the lift for different *x*-values. As Verbeke et al. (2012) mention, this allows to draw more correct conclusions compared to a single lift number when comparing models. While both AUC and lift are generally accepted for evaluating data mining models, the current setting favors lift over AUC. Indeed, the company can only contact a limited top-fraction of prospects within budget limit (typically, also for CCR, 5-10% of the prospects). Hence, we want the model that best identifies the top-fraction of prospects relevant to the company as given by the lift, not necessarily the best overall model. All results show the median AUC and lift curve of the median model of our 5*2 CV procedure. We evaluate whether AUC values are statistically significant using the Wilcoxon signed-rank test (Demšar, 2006).

4. Results

We summarize the results for the different models using AUC and lift. The AUCs are given in Table 4.2, while the lift curves are plotted in Figures 4.2-4.3. The results show that the AUCs range from 0.534 to 0.590 for the Phase I model, and from 0.537 to 0.612 for the Phase II model. These values are all significantly different from the random model with AUC = 0.5 (V = 55, p =< 0.001). These AUC values are not impressive when compared to for instance reported AUC values of churn prediction models (e.g., Larivière and Van den Poel, 2005). However, they are comparable to the results found in acquisition literature (e.g., D'Haen et al., 2016; Thorleuchter et al., 2012, with maximum AUC values of 0.62 and 0.61 respectively)[4], which demonstrates the difficulty of acquisition prediction. For managerial recommendations, lift is more useful

[4] Note that the results of e.g. (D'Haen et al., 2013) are not directly comparable because they evaluate with current customers and prospects, instead of contacting prospects and evaluating conversion. When applying the same technique here, the AUC varies between 0.69 and 0.75.

and we note that the best performing models have acceptable lift curves comparable to previous literature (e.g., Thorleuchter et al. (2012) reports top-10% lift ranging from 1.35 to 1.65).

By investigating model M1, M2 and M3, we can derive the value of Facebook pages compared to website and commercial information. With regard to the Phase I models' AUCs, we find that Facebook data is significantly better than commercial data ($V = 55$, $p < 0.001$), but only slightly better than website data ($V = 43$, $p = 0.07$). These results are confirmed in terms of the lift curves (Figure 4.2) although the lift curve for the website model is slightly higher than the Facebook model for the top 5% lift. Phase II models show that Facebook data is better compared to both commercial data and website data ($V = 55$, $p < 0.001$ in both cases). This is confirmed by Figure 4.3, which indicates the higher power of Facebook data for Phase II in terms of lift. The upper four lines are models that include Facebook variables while the lower three lines do not, which indicates that Facebook pages perform clearly better for the Phase II models.

Table 4.2: (median) AUC of all models

Data	Phase I	Phase II
Commercial (M1)	0.534	*0.554*
Website (M2)	*0.556*	0.537
Facebook (M3)	0.565	*0.607*
Commercial + website (M4)	*0.581*	0.561
Commercial + Facebook (M5)	0.584	***0.612***
Website + Facebook (M6)	0.584	*0.606*
Commercial + Facebook + website (M7)	**0.590**	*0.607*

Bold values indicate the highest performance of AUC per phase; Italic values indicate the highest value of AUC per model.

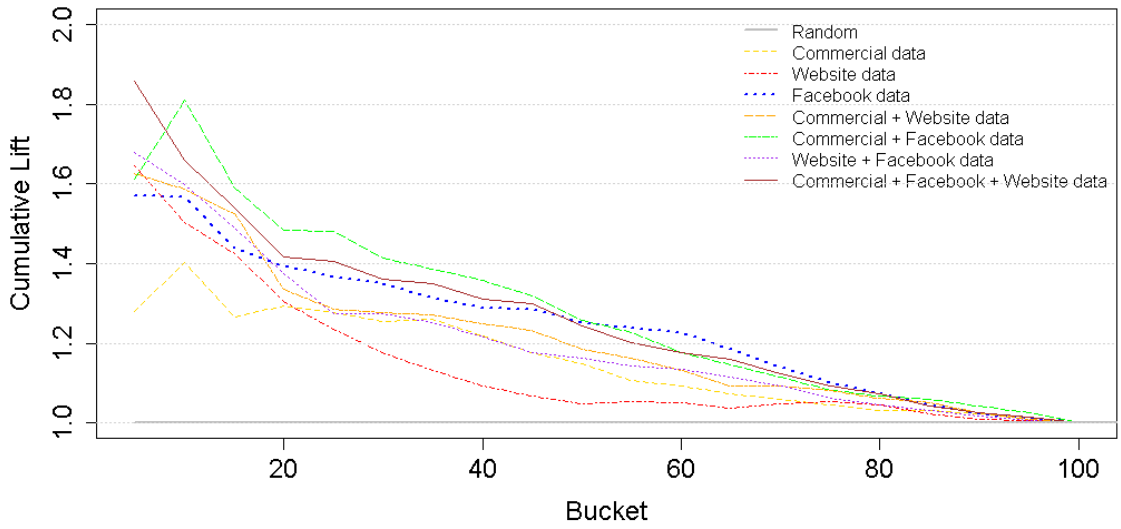


Figure 4.2: (median) Cumulative lift curve for Phase I model

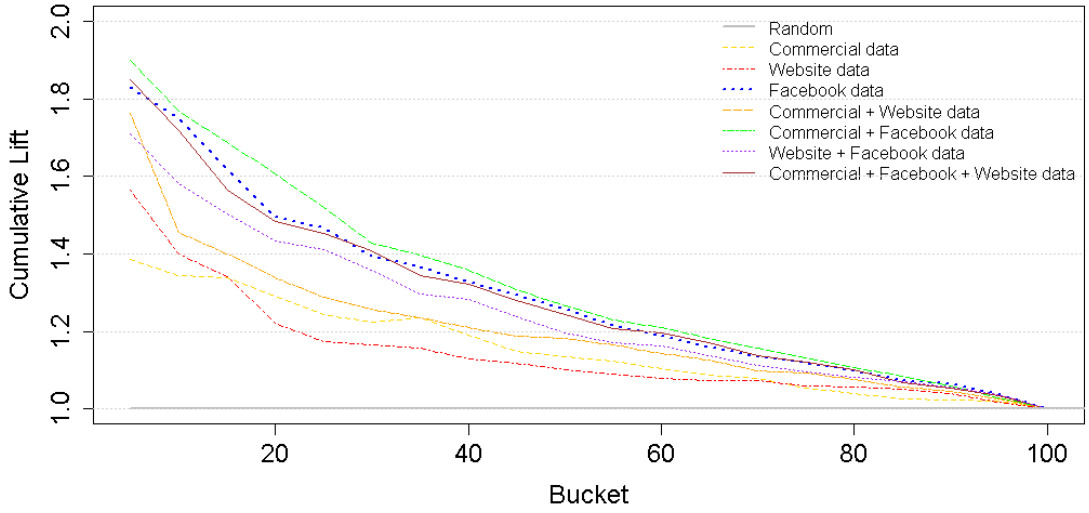


Figure 4.3: Cumulative lift curve for Phase II model

Next, we evaluate how models perform when they combine the different data sources. First, with regard to the added value of freely available data over commercial data, we compare model M4 and M5 with M1. This shows that both AUCs and lift curves indicate superior performance of combined models in both Phases (all $V = 55$, $p < 0.001$ except for M4 in Phase II: $V = 42$, $p = 0.08$). When we combine the two freely available data sources (M6), we see that this performs better compared to the single sources in Phase I (although not significantly). However, the performance deteriorates compared to Facebook data in Phase II ($V = 52$, $p = 0.005$) due to the bad performance of the website data in this Phase. Moreover, the lift curves

in Phase I and II show that the lift curves of M6 are pulled down due to the bad lift curves of the website model.

The performance of models that combine all three data sources is also more ambiguous (M7 compared to M4, M5 or M6). The Phase I model indicates better performance of the most complete model M7 in terms of AUC (not significant), but the complete lift curves are more in favor of M5. The Phase II model indicates that combining all three datasets gives worse prediction performance compared to M5, both in terms of AUC and lift ($V = 51$, $p = 0.007$).

Finally, we want to compare the results of our Phase II models with the results of the Phase I models. The AUC results show that in five of the seven models, the Phase II models perform better compared to the Phase I models, with a striking difference in model M3. In model M2 and M4, the performance of the Phase II model is lower compared to the Phase I model. These models contain the website information. An explanation behind this is that the analysis of unstructured data, such as website data, is very dependent on the amount of information in the training set. Martens et al. (2016) show that an increased amount of training data, especially for unstructured information, results in higher AUC. Thus, given the textual, unstructured nature of website information, it is likely that the same behavior applies. In the Phase I model, the training data consists of both customers and prospects in the training data ($n \approx 13250$), while Phase II only uses the prospects in the training data ($n \approx 4500$). However, Phase II can be retrained every time new information becomes available (new feedback from the call center actions), which would increase the size of the training set in future runs.

The results in terms of lift are somewhat more ambiguous, but in general support the results in terms of AUC.

5. Discussion

Our results show that the sales process can be improved by using social media in a way that was not explored yet, i.e. using a data mining approach to social media in a formal information system. Within this research, we have shown that automatic handling of Facebook pages is a valuable tool to (1) improve the qualification prediction of prospects into leads worth pursuing and (2) reduce the time needed to screen the Facebook pages drastically. We believe that an information system based on this new approach is capable of making the sales process more efficient, at least for companies with standardized products and with a relatively simple sales process meant to serve a lot of prospects (Homburg et al., 2011). We will discuss several key insights in the following paragraphs.

5.1. Added Value of Facebook Information

We have used Facebook pages of prospects instead of personal profiles of prospective customers' salespersons, as was common in literature (e.g., Andzulis et al. (2012); Giamanco and Gregoire (2012); Marshall et al. (2012)). We hypothesize social media pages to be the most informative, and our models generally support these expectations. Moreover, we argued that it was mainly the combination of company characteristics and attitudinal information that makes social media a strong predictor.

We further investigate this statement by modeling the company characteristics and attitudinal information separately in a Random Forest model (we take the median model of the 5*2 fold CV Phase II model for this extra analysis). The two models have similar performance, yielding an AUC of 0.567 for the company characteristics and 0.551 for the attitudinal information. We can state that both sources of information are valuable for the prediction exercise. Moreover, we see that the two data types are complementary. We show this by evaluating the AUC of the complete Facebook model (0.607). Its added value over a random model (AUC = 0.5) is 0.107. For the individual models, the added values over a random model are 0.067 and 0.051 respectively, summing up to a total added value of 0.118. The ratio of both added values is 0.907 (0.107/0.118), which indicates that 90% of the added value of the both individual models is retained in the complete model. This proves that the two data types within Facebook data contain different information, which renders the Facebook pages the most interesting data source.

We can also conclude that the company characteristics contained within the Facebook information do a slightly better job in predicting successful conversions compared to characteristics in commercial information (with an AUC of 0.554, see Table 2). Moreover, the combination of the two data sources shows an increase over the two individual data sources, indicating that there is complementary information in the two data sets. Thus, the company characteristics in the two dataset are not entirely the same, yielding additional insights for the prediction exercise.

Finally, we show the added value of the Facebook variables by looking at the variable importance plots generated from the Random Forest models in Figure 4.4 and 4.5. These plots show the 50 most important variables for each Phase of the most complete model, and we labeled the top 10. In both cases, all top 10 variables are Facebook variables. These plots show that the number of Likes, Check-ins and Were-Here were most influential in both models.

Interestingly, they also show that none of the commercial data variables is among the top-50 variables in the Phase I most complete model.

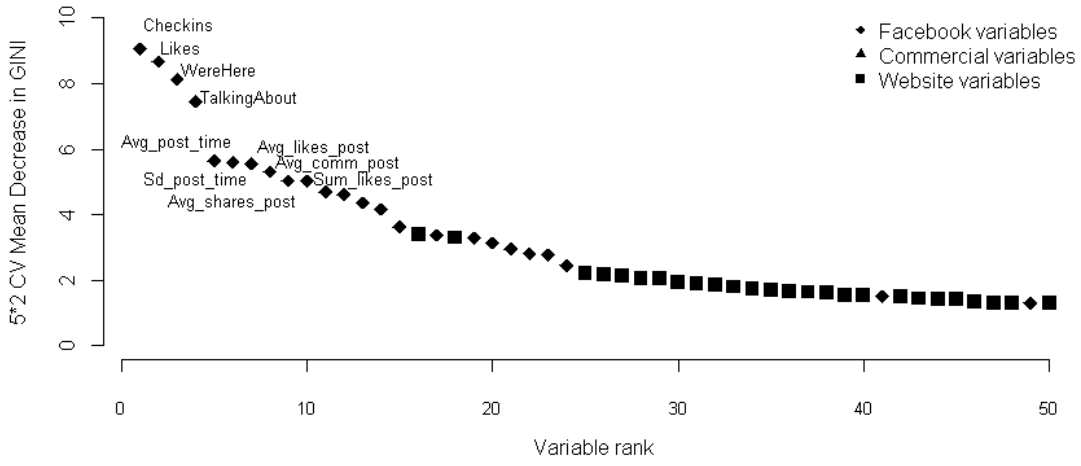


Figure 4.4: (median) variable importance plot for Phase I model 7

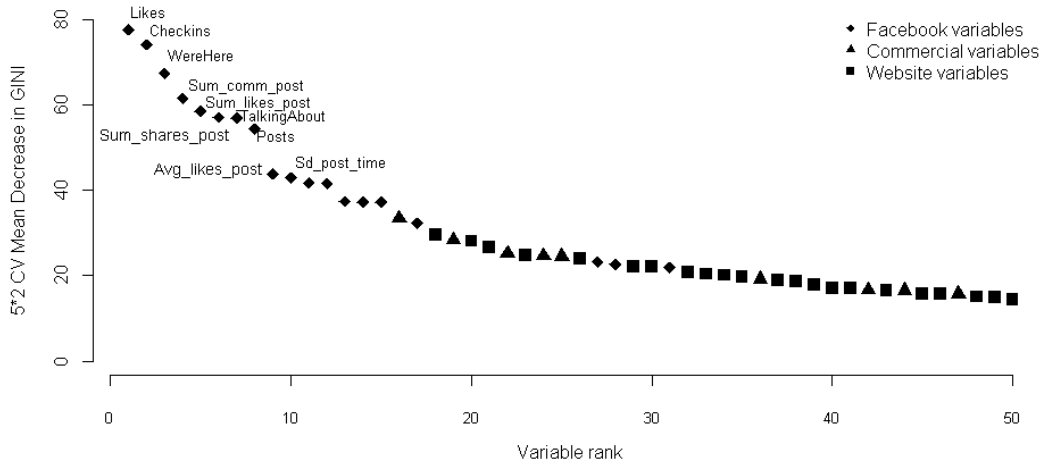


Figure 4.5: (median) variable importance plot for Phase II model 7

5.2. Combination of Data Sources

Based on previous studies and previous work in marketing and data mining (D'Haen et al., 2013; Goel and Goldstein, 2013; Hanssens et al., 2014), one would expect a combination of data sources to outperform models that are based on a single data source. This is only partially true for our models. For the Phase I models, combining all data sources indeed yielded best performance in

terms of AUC, while the lift curves were more in favor of a model combining Facebook data and commercial data. For the Phase II models, the model combining Facebook data and commercial data proved to be the best. Another important conclusion is that it may be worth to build models based on freely available data only. Commercial data, sold by specialized vendors, come at large costs which may not be worth the relatively small increase in model performance. Indeed, the models that use Facebook data and/or website data perform almost equally well (the combined Facebook and web model for Phase I, the Facebook only model in Phase II). Thus, it can be an interesting exercise to trade off higher model performance (and conversion ratio) with higher costs of data collection to optimize budget spending. Note that in our research, we also use commercial data to identify prospects and their websites. We believe, however, that social media (e.g., online review sites or Facebook) now provide opportunities to search online for potential prospects, although developing this search models again requires effort and time. When assessing the trade-off with commercial data, one thus also needs to take account the extra effort it takes to construct a prospect list when no commercial data are available.

5.3. Iterative Process of the Sales Funnel Model

The results show that the Phase II model performs better compared to the Phase I model, which is in line with expectations. Indeed, the goal of the study and sales process is to separate good from bad prospects. In Phase II, we explicitly model good or converted prospects. In Phase I, we aim to identify good prospects by comparing them to existing customers. However, as noted by Blattberg et al. (2008), these Phase I models are not necessarily very predictive of which prospects will actually purchase. The framework of D'Haen and Van den Poel (2013) further suggests that the process is iterative, because new feedback can be fed into the model as time goes by, increasing the amount of data available for training the model. This offers potential to increase the relatively low performance of the acquisition models, as Martens et al. (2016) showed that an increased training sample can increase AUC. Lilien (2016) mentions that practitioners are starting to use look-alike models to qualify prospects. We encourage to go further and adopt the phased model to increase performance even more.

5.4. Practical Implications

Finally, we show the economic value of our models by calculating the monetary savings that can be achieved (Hanssens and Pauwels, 2016). We adapt the churn profitability analysis in Neslin et al. (2006) for the acquisition case, and define the financial gains of an acquisition campaign as a function of the ability of the predictive model to identify would-be customers

$$\Pi = N\alpha\,[\beta(R - c - S) - c\,(1 - \beta)], \tag{4.3}$$

where Π is the financial gain of the campaign, N is the total number of prospects, α is the fraction of prospects targeted, β is the percentage of prospects that could be converted to customers, R is the average one-year revenue of a new customer for CCR, c is the contacting cost per prospect and S is the cost of a salesperson for closing the deal. Note that we are using revenue instead of profits for reasons of confidentiality. The first term between brackets reflects the contribution of the converted prospects, while the latter term reflects the cost of contacting non-converted prospects. As in Neslin et al. (2006), β reflects the model's accuracy and can be expressed as the multiplication of β_0 and λ,

$$\beta = \lambda\,\beta_0, \tag{4.4}$$

where β_0 represents the overall prospect to customer conversion rate and λ is the lift of the model. For a random calling model, we expect average performance and λ = 1. Substituting Equation 4 in Equation 3 yields:

$$\Pi = N\alpha\,[\lambda\,\beta_0(R - c - S) - c\,(1 - \lambda\,\beta_0)]. \tag{4.5}$$

Simplifying this equation leads to:

$$\Pi = N\alpha\,[\lambda\,\beta_0(R - S) - c\,]. \tag{4.6}$$

We analyze the financial gains for each of our Phase II models, which is summarized in Table 3 CCR has approximately one million prospects, so we take N to be one million. Assume CCR calls the top 5% prospects (α = 0.05), which means we use the top 5%-lift as λ[5], given by the first column in Table 3. The results of the real-life experiment show that β_0 is equal to 7.4%. R, the average one-year revenue per new customer, is calculated to be approximately \$8,000 based on previous converted prospects. Finally, we assume the cost of contacting a prospect, c, to be \$50 and the cost of a salesperson for closing the deal, S, to be \$500.

[5] Note that, as Verbeke et al. (2012) correctly points out, the regularly chosen values for α of 5 or 10% are not necessarily the most optimal values in terms of return, and that these values do not need to be the same among all models. However, in our case, the difference between the potential revenue and costs is so large that even random calling is still profitable (which is also the current situation). Therefore, we chose a realistic percentage that the company is able to contact and which is in line with their current practice.

Table 4.3: Financial gains for Phase II models

Data	Top 5% Lift	Top 5% Response ($= \lambda \beta_0$)	Financial gain ($, in millions)
Baseline	1	7.4%	27.100
Commercial (M1)	1.385	10.2%	35.934
Website (M2)	1.566	11.6%	40.957
Facebook (M3)	1.830	13.5%	48.283
Commercial + website (M4)	1.901	14.1%	50.253
Commercial + Facebook (M5)	1.763	13.0%	46.423
Website + Facebook (M6)	1.710	12.7%	44.953
Commercial + Facebook + website (M7)	1.850	13.7%	48.838

Currently, the company is not using any model to select the most interesting prospects, although they have commercial data available (the baseline in Table 4.3). Table 4.3 shows that even for the worst model, M1, an increased response percentage of 2.8%-points (10.2% - 7.4%) can be achieved, which is equal to 1,425 (2.8% * 50K) extra customers that are likely to be converted, or additional financial gains of $8,834,000. For the best model, M4, there was an increase from 7.4% to 14.1%, an increase of 6.7%-points. This implies that with the best model, more than 3330 extra customers can be generated without extra sales cost, resulting in increased financial benefits of $21,738,000. This clearly shows the usefulness and value of the model and the Facebook dataset in particular. Next to the financial gains, we also note that more subtle gains may be achieved by using a formal information system. Indeed, by automatically collecting, cleaning and analyzing the Facebook pages, the sales representatives' time can be spent more efficiently.

6. Conclusions and future research

This paper assesses the added value of social media pages in a Business-to-Business customer acquisition system, taking a big data view on social media. More specifically, we evaluate the predictive value of data extracted from the prospects' Facebook pages in a customer acquisition context. We test our approach using a real-life field study at Coca Cola

Refreshments USA. The results show that models including Facebook data are substantially better at predicting 'good' prospects. Moreover, the results show that Facebook data and the other data sources contain complementary information.

With this research, we answer recent calls (e.g. Lilien, 2016) for research on B2B customer analytics, as this is heavily under-researched compared to B2C customer analytics. From this point of view, we are, to the best of our knowledge, the first to investigate the added value of social media within a B2B context quantitatively. We show that the richness of social media has value in discovering good prospects. From the point of view of B2B acquisition modeling, we provide evidence that new data sources such as social media can and should be used to further improve the predictive performance. Moreover, we show on a large scale that the modeling exercise can be improved by taking into account the iterative nature of the sales process.

Finally, we want to consider several limitations of this study which could prove important for future extensions of this work. First, although the sales funnel is presented as a simple process, it might be complex in reality. While the seller can try to qualify prospects, prospects' propensity to purchase is also driven by various actions (Kumar et al., 2013). These include (1) seller initiated efforts, (2) competitor initiated efforts, (3) client initiated efforts and (4) prospect characteristics (Kumar et al., 2013; Reinartz et al., 2005). Within this study, we will limit ourselves mainly to the inclusion of prospect characteristics for qualifying prospects, and seller initiated efforts for contacting the prospect. The two other actions are difficult to measure in the prospect stage of the sales funnel, certainly at a large scale.

Second, the prediction models focus on specific samples, that were identified by commercial data and had both a website and Facebook page available. This possibly leads to selection bias, as the behavior of prospects that do not have a website or Facebook page available may be fundamentally different compared to the ones who do. For example, bars and restaurants with relatively older operators may be less likely to be active on social media. At the same time, they may have a lower propensity to become client of CCR because their clientele is not that much interested in soft drinks. Since we apply the models within a prediction context to similar samples, and we do not aim to extract managerial recommendations on specific variables or actions, this selection bias does not harm the analyses. If we would look for variables to act upon, we would suggest to use Heckman selection models. Finally, we want to mention that for customers not in our sample, simpler models could be built based on for

instance commercially available data. Taking this subset of data (outlets with only commercial data) yielded similar performance as the M1 model used (AUC between 0.54 - 0.55).

Third, uplift modeling could be adopted as an alternative to our classic predictive approach. The classic models output a response probability that the prospect company will buy, maybe after some campaign (e.g., marketing initiatives, calls, visits) from the company looking for customers. Uplift modeling states that one should not estimate the response probabilities, but the change in response probabilities caused by the campaign (in comparison to no campaign) (Kane et al., 2014) in order to target only those prospects most influenced by the campaign. Therefore, in uplift modeling, control and treatment groups are set up to measure the 'true lift'. In this study, a part of the prospects would not be contacted, and the conversion rate of the contacted prospects versus the rate of the non-contacted prospects for the entire lift curve should be evaluated.

Finally, future work might assess the added value of social media pages for profitability prediction instead of prospect conversion (Reinartz et al., 2005). When a longer timeframe becomes available (e.g., after 1 year), the profitability of the converted prospects can be assessed. Subsequently, a two-stage model can be built to predict not only which prospects will convert, but also which of those converted prospects are more likely to become profitable customers for the company in the near future. As such, the sales process would become even more effective by not spending resources on unprofitable prospects.

7. References

Ahearne, M., Jones, E., Rapp, A., Mathieu, J., 2008. High Touch Through High Tech: The Impact of Salesperson Technology Usage on Sales Performance via Mediating Mechanisms. Management Science 54, 671–685.

Andzulis, J. "Mick," Panagopoulos, N.G., Rapp, A., 2012. A Review of Social Media and Implications for the Sales Process. Journal of Personal Selling & Sales Management 32, 305–316.

Baesens, B., Bapna, R., Marsden, J., Vanthienen, J., Zhao, J., 2016. Transformational issues of big data and analytics in networked business. MIS Quarterly 40, 807–818.

Blattberg, R.C., Kim, B.-D., Neslin, S.A., 2008. Database Marketing, International Series in Quantitative Marketing. Springer New York, New York, NY.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Brito Pereira Zamith, E., Zanette, M.C., Caires Abdalla, C., Ferreira, M., Limongi, R., Rosenthal, B., 2015. Corporate Branding in Facebook Fan Pages: Ideas for Improving Your Brand Value, Digital and Social Media Marketing and Advertising Collection. Business Expert Press.

Chen, D.Q., Preston, D.S., Swink, M., 2015. How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management. Journal of Management Information Systems 32, 4–39.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41, 391–407.

Demšar, J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1–30.

D'Haen, J., Van den Poel, D., 2013. Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. Industrial Marketing Management, Special Issue on Applied Intelligent Systems in Business-to-Business Marketing 42, 544–551.

D'Haen, J., Van den Poel, D., Thorleuchter, D., 2013. Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. Expert Systems with Applications 40, 2007–2012.

D'Haen, J., Van den Poel, D., Thorleuchter, D., Benoit, D.F., 2016. Integrating expert knowledge and multilingual web crawling data in a lead qualification system. Decision Support Systems 82, 69–78.

Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10, 1895–1923.

Facebook, 2016. Company Info | Facebook Newsroom.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15, 3133–3181.

Giamanco, B., Gregoire, K., 2012. Tweet Me, Friend Me, Make Me Buy. Harvard Business Review 90, 88–93.

Goel, S., Goldstein, D.G., 2013. Predicting Individual Behavior with Social Networks. Marketing Science 33, 82–93.

Goh, K.-Y., Heng, C.-S., Lin, Z., 2013. Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research 24, 88–107.

Hansotia, B.J., Wang, P., 1997. Analytical challenges in customer acquisition. Journal of Interactive Marketing 11, 7–19.

Hanssens, D.M., Pauwels, K.H., 2016. Demonstrating the Value of Marketing. Journal of Marketing 80, 173–190.

Hanssens, D.M., Pauwels, K.H., Srinivasan, S., Vanhuele, M., Yildirim, G., 2014. Consumer Attitude Metrics for Guiding Marketing Mix Decisions. Marketing Science 33, 534–550.

Homburg, C., Müller, M., Klarmann, M., 2011. When Should the Customer Really Be King? On the Optimum Level of Salesperson Customer Orientation in Sales Encounters. Journal of Marketing 75, 55–74.

Järvinen, J., Taiminen, H., 2016. Harnessing marketing automation for B2B content marketing. Industrial Marketing Management 54, 164–175.

Jolson, M.A., 1988. Qualifying sales leads: The tight and loose approaches. Industrial Marketing Management 17, 189–196.

Kane, K., Lo, V.S.Y., Zheng, J., 2014. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analysis 2, 218–238.

Keller, K.L., Richey, K., 2006. The importance of corporate brand personality traits to a successful 21st century business. Journal of Brand Management 14, 74–81.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., Kannan, P. k., 2015. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. Journal of Marketing 80, 7–25.

Kumar, V., Petersen, J.A., Leone, R.P., 2013. Defining, Measuring, and Managing Business Reference Value. Journal of Marketing 77, 68–86.

Laiderman, J., 2005. A structured approach to B2B segmentation. Journal of Database Marketing and Customer Strategy Management 13, 64–75.

Larivière, B., Van den Poel, D., 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29, 472–484.

Lash, M.T., Zhao, K., 2016. Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems 33, 874–903.
Leverage New Age Media, 2015. Social Media Comparison Infographic. Leverage New Age Media.
Lilien, G.L., 2016. The B2B Knowledge Gap. International Journal of Research in Marketing 33, 543–556.
Lix, T.S., Berger, P.D., Magliozzi, T.L., 1995. New customer acquisition: prospecting models and the use of commercially available external data. Journal of Direct Marketing 9, 8–18.
Long, M.M., Tellefsen, T., Lichtenthal, J.D., 2007. Internet integration into the industrial selling process: A step-by-step approach. Industrial Marketing Management 36, 676–689.
Marshall, G.W., Moncrief, W.C., Rudd, J.M., Lee, N., 2012. Revolution in Sales: The Impact of Social Media and Related Technology on the Selling Environment. Journal of Personal Selling & Sales Management 32, 349–363.
Martens, D., Provost, F., Clark, J., Junqué de Forunty, E., 2016. Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. MIS Quarterly 40, 869–888.
Meire, M., Ballings, M., Van den Poel, D., 2016. The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems 89, 98–112.
Melville, P., Rosset, S., Lawrence, R.D., 2008. Customer Targeting Models Using Actively-selected Web Content, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08. ACM, New York, NY, USA, pp. 946–953.
Michaelidou, N., Siamagka, N.-T., Christodoulides, G., 2011. Usage, barriers and measurement of social media marketing: an exploratory investigation of small and medium B2B brands. Industrial Marketing Management 40, 1153–1159.
Monat, J.P., 2011. Industrial sales lead conversion modeling. Marketing Intelligence & Planning 29, 178–194.
MSI Research Priorities 2016-2018, 2016. New data, new methods, and new skills — how to bring it all together? [WWW Document]. Marketing Science Institute.
Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., Mason, C.H., 2006. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research 43, 204–211.
Reinartz, W., Thomas, J.S., Kumar, V., 2005. Balancing Acquisition and Retention Resources to Maximize Customer Profitability. Journal of Marketing 69, 63–79.
Reinartz, W.J., Kumar, V., 2003. The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing 67, 77–99.
Rodriguez, M., Peterson, R.M., Krishnan, V., 2012. Social Media's Influence on Business-to-Business Sales Performance. Journal of Personal Selling & Sales Management 32, 365–378.
Sabnis, G., Chatterjee, S.C., Grewal, R., Lilien, G.L., 2013. The Sales Lead Black Hole: On Sales Reps' Follow-Up of Marketing Leads. Journal of Marketing 77, 52–67.
Schillewaert, N., Ahearne, M.J., Frambach, R.T., Moenaert, R.K., 2005. The adoption of information technology in the sales force. Industrial Marketing Management, Technology and the Sales Force 34, 323–336.
Schwartz, E.M., Bradlow, E.T., Fader, P.S., 2014. Model Selection Using Database Characteristics: Developing a Classification Tree for Longitudinal Incidence Data. Marketing Science 33, 188–205.
Thorleuchter, D., Van den Poel, D., Prinzie, A., 2012. Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Systems with Applications 39, 2597–2605.
Trainor, K.J., Andzulis, J. (Mick), Rapp, A., Agnihotri, R., 2014. Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. Journal of Business Research 67, 1201–1208.
van Wangenheim, F., Bayón, T., 2007. The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition. Journal of the Academy of Marketing Science 35, 233–249.

VentureBeat, 2016. Facebook: 60 million businesses have Pages, 4 million actively advertise [WWW Document]. VentureBeat. URL http://venturebeat.com/2016/09/27/facebook-60-million-businesses-have-pages-4-million-actively-advertise/ (accessed 1.5.17).

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218, 211–229.

Verbeke, W.J., Belschak, F.D., Bakker, A.B., Dietz, B., 2008. When Intelligence Is (Dys)Functional for Achieving Sales Performance. Journal of Marketing 72, 44–57.

Walker, O.C., Churchill, G.A., Ford, N.M., 1977. Motivation and Performance in Industrial Selling: Present Knowledge and Needed Research. Journal of Marketing Research 14, 156–168.

Webster, F.E., Wind, Y., 1972. A General Model for Understanding Organizational Buying Behavior. Journal of Marketing 36, 12–19.

Wedel, M., Kannan, P. k., 2016. Marketing Analytics for Data-Rich Environments. Journal of Marketing 80, 97–121.

Weitz, B.A., Sujan, H., Sujan, M., 1986. Knowledge, Motivation, and Adaptive Behavior: A Framework for Improving Selling Effectiveness. Journal of Marketing 50, 174–191.

Xie, K., Lee, Y.-J., 2015. Social Media and Brand Purchase: Quantifying the Effects of Exposures to Earned and Owned Social Media Activities in a Two-Stage Decision Making Model. Journal of Management Information Systems 32, 204–238.

Yu, Y., Cai, S., 2007. A new approach to customer targeting under conditions of information shortage. Marketing Intelligence & Planning 25, 343–359.

Zoltners, A.A., Sinha, P., Lorimer, S.E., 2008. Sales Force Effectiveness: A Framework for Researchers and Practitioners. Journal of Personal Selling & Sales Management 28, 115–131.

8. Appendix

Appendix A: Variable description

Commercial Data

Variable name	*Variable description*	*Proportion (range)*
Contact	Dummy indicating whether contact info was present	0.727
Fax	Dummy indicating whether fax number was available	0.261
City_1 – City_7	Dummy for city (7 dummies for the largest cities which represent 10% of the database)	0.001-0.027
State_1 – State_7	Dummy for state (7 dummies for the largest states, which represent 56% of the database)	0.042-0.171
Region_1 – Region_2	Dummy indicating region (2 dummies)	0.400-0.329
Tz_1 – Tz_4	Dummy indicating time zone (4 time zone dummies)	0.006-0.480
Ind_1 – Ind_7	Dummy indicating Industry code (7 dummies)	0.030-0.479
Type_1 – Type_7	Dummy indicating type of outlet (7 dummies)	0.018-0.700
Emp_A – Emp_F	Dummy indicating employee size (6 dummies, ranging from A (1-4 employees) to F (100-500> 500 employees))	0.001-0.339
Rev_A – Rev_F	Dummy indicating the annual revenue estimation (6 dummies, ranging from A (< $ 500,000) to F ($ 10-20 million))	0.002-0.494
Ad_1 – Ad_4	Dummy indicating the Ad Size (4 dummies)	0.710-0.013
SqFt_1	Dummy indicating square footage (only 2 types were available)	0.604
CS_1 – CS_5	Dummy indicating the credit score (5 dummies)	0.005-0.104
Gender	Gender of the outlet owner (2 dummies, since missing is included as category)	0.187-0.471

Language	Language spoken by the outlet owner (3 dummies)	0.052-0.471
Ethnic code	Ethnic group of the outlet owner (7 dummies)	0.001-0.243
Census_gender	Average male proportion in the neighborhood according to the census	0.4966
Census_HHNbr	Number of households in the neighborhood according to the census	
Census_HHIncome	Income of households in the neighborhood according to the census	

Website Data

Variable name	*Variable description*	*Proportion (range)*
Concept1 – Concept50	50 concepts obtained via LSI	
Facebook	Dummy indicating whether the website indicated Facebook presence	0.346
Twitter	Dummy indicating whether the website indicated Twitter presence	0.212
Instagram	Dummy indicating whether the website indicated Instagram presence	0.095

Facebook data

Variable name	*Variable description*	*Proportion (range)*
Website	Dummy indicating presence of website on Facebook page	0.880
Phone	Dummy indicating presence of phone number on Facebook page	0.875
Location	Dummy indicating presence of location on Facebook page	0.851
Description	Dummy indicating presence of a description of the outlet on Facebook page	0.536
About	Dummy indicating presence of the 'about' section on Facebook page	0.779
Price_1 – Price_4	Dummy indicating the price range (4 dummies , < $ 10 to > $50)	0.010-0.295
Cat_1 – Cat_50	Dummy indicating type of outlet on Facebook (50 dummies)	0.003-0.062
Delivery	Dummy indicating whether there is delivery service	0.121
Catering	Dummy indicating whether there is catering service	0.348
Group	Dummy indicating whether there is a possible group service	0.457
Kids	Dummy indicating whether there is kids service	0.397
Outdoor	Dummy indicating whether there is outdoor service	0.267
Reservation	Dummy indicating whether there is reservation service	0.300
Takeout	Dummy indicating whether there is takeout possibility	0.494
Waiter	Dummy indicating whether there is a waiter	0.360
Walk-in	Dummy indicating whether there is a walk-in service	0.504
Breakfast	Dummy indicating whether there is possibility for breakfast	0.160
Coffee	Dummy indicating whether there is possibility for coffee	0.191
Dinner	Dummy indicating whether there is possibility for dinner	0.502
Drinks	Dummy indicating whether there is possibility for drinks	0.341
Lunch	Dummy indicating whether there is possibility for lunch	0.487
Parking_1 – Parking_3	Dummy indicating parking availability (3 dummies)	0.042-0.464
Community	Dummy indicating whether the page is a (non-official) community page	0.092
Likes[1]	The number of likes the page has received	
Checkins[1]	The number of check-ins the page has received	
WereHere[1]	The sum of all people indicating their presence at the outlet	
TalkingAbout[1]	Total number of people talking about the outlet	
Avg_likes_post[2]	The average number of likes on posts of the outlet	
Avg_comm_post[2]	The average number of comments on posts of the outlet	
Avg_shares_post[2]	The average number of shares on posts of the outlet	
Posts[2]	The total number of Facebook posts of the outlet	
Sum_likes_post[2]	The total number of likes on posts of the outlet	
Sum_comm_post[2]	The total of comments on posts of the outlet	

Sum_share_post[2]	The total number of shares on posts of the outlet
Avg_post_time[2]	Average time between two Facebook posts of the outlet
Sd_post_time[2]	Standard deviation of the time between two Facebook posts of the outlet
Avg_likes_tagpost[2]	Average number of likes on posts in which the outlet was tagged
Avg_comm_tagpost[2]	Average number of comments on posts in which the outlet was tagged
Avg_shares_tagpost[2]	Average number of shares on posts in which the outlet was tagged
Tagged_posts[2]	The number of posts in which the outlet was tagged
Sum_likes_tagpost[2]	The total number of likes on posts in which the outlet was tagged
Sum_comm_tagpost[2]	The total number of comments on posts in which the outlet was tagged
Sum_shares_tagpost[2]	The total number of shares of posts in which the outlet was tagged
Avg_tagpost_time[2]	Average time between two Facebook posts in which the outlet was tagged
Sd_tagpost_time[2]	Standard deviation of the time between two Facebook posts in which the outlet was tagged

[1] Number at the time of scraping
[2] Number over six months prior to scraping

Appendix B: Descriptive statistics of continuous variables

Variable	Min	Max	Mean	Median	Standard deviation
Census_HHNbr	0	714	237	234	61
Census_HHIncome	0	632.945	77.581	67.273	45.452
Likes	0	170 743 168	775 615	1 631	4 332 566
Checkins	0	25 001 314	11 807	1 469	200 198
WereHere	0	28 532 271	205 112	1 668	1 064 094
TalkingAbout	0	1 401 564	3 296	32	24 682
Avg_likes_post	0	230 255	1 416	5	8 199
Avg_comm_post	0	6 844	27	0	143
Avg_shares_post	0	15 815	42	2	282
Posts	0	18.780	926	34	3.180
Sum_likes_post	0	940.442.372	22.836.123	222	142.714.920
Sum_comments_post	0	14.269.655	361.471	15	2.167.855
Sum_share_post	0	16.666.602	471.394	19	2.634.769
Avg_post_time	0	166	3	1	9
Sd_post_time	0	148	8	3	12
Avg_likes_tagpost	0	14.818	5	0	101
Avg_comm_tagpost	0	4.174	1	0	26
Avg_shares_tagpost	0	1.133	2	0	14
Tagged_posts	0	62.700	1.865	1	9.604
Sum_likes_tagpost	0	1.338.016	34.297	0	203.466
Sum_comm_tagpost	0	303.468	7.759	0	46.104
Sum_shares_tagpost	0	10.051	60	0	411
Avg_tagpost_time	0	177	11	15	13
Sd_tagpost_time	0	163	11	15	11

5

General Discussion

Social media have changed the way customers and business interact. On the one hand, customers (or even prospects, ex-customers, and complete strangers) can formulate their opinion about brands, products or services on social media and hence influence other (potential) customers. As such, it is threatening existing business models. On the other hand, it offers businesses a new, interactive way to reach out to customers, to foster engagement, and thus creates new opportunities for businesses (Hennig-Thurau et al., 2010). Companies should thus adapt to the changing environment, and try to understand and use social media as a part of the communication and marketing mix (Chen and Xie, 2008). While many companies have already adopted social media, academic research regarding social media is still relatively scarce. Viral marketing campaigns have been well researched (e.g., Hinz et al., 2011; van der Lans et al., 2010, Zhang et al., 2017), and the value of user and marketer generated content on social media is also subject of a growing amount of literature (e.g., Babić Rosario et al., 2016). Recently, the value of a Facebook 'like' has gotten some attention as well. However, in the light of the importance and abundance of social media nowadays, the research is limited. For instance, research on the individual customer level is scarce, as well as research that explains if and how to use social media (quantitatively) in a Business-to-Business context. Therefore, this dissertation wants to add to the literature by arguing that social media can have value for businesses in many different ways.

In the remainder of this chapter, we first rehearse the outlook of this dissertation and the linkage between the chapters. Next, we provide a brief summary of each of the chapters,

followed by a discussion of the contributions. Finally, we provide an outlook for future research and potential difficulties with research in social media.

1. Outlook of the dissertation

The different chapters of this dissertation are visually represented in Figure 5.1 (retake of Figure 1.1). We analyzed the business value of social media from different perspectives. First, it is important to note that we focus on the largest social network, Facebook. Facebook contains user profiles (linked to customers) and fan pages (linked to companies). In this dissertation, we focus on both user profiles (chapter 2) and fan pages (chapter 4), and also on the link between user profile and fan pages (3). Chapter 2 details how firms can improve customer sentiment prediction of customer Facebook posts. Chapter 3 evaluates how customers' sentiment (expressed by Facebook comments) related to actual experience encounters of a soccer team can be moderated by MGC, and linked to customer lifetime value. Finally, chapter 4 evaluates how Facebook fan pages of prospect companies can be used for prospect to customer conversion. All in all, this dissertation provides a comprehensive view of the value of Facebook as business tool over the different chapters

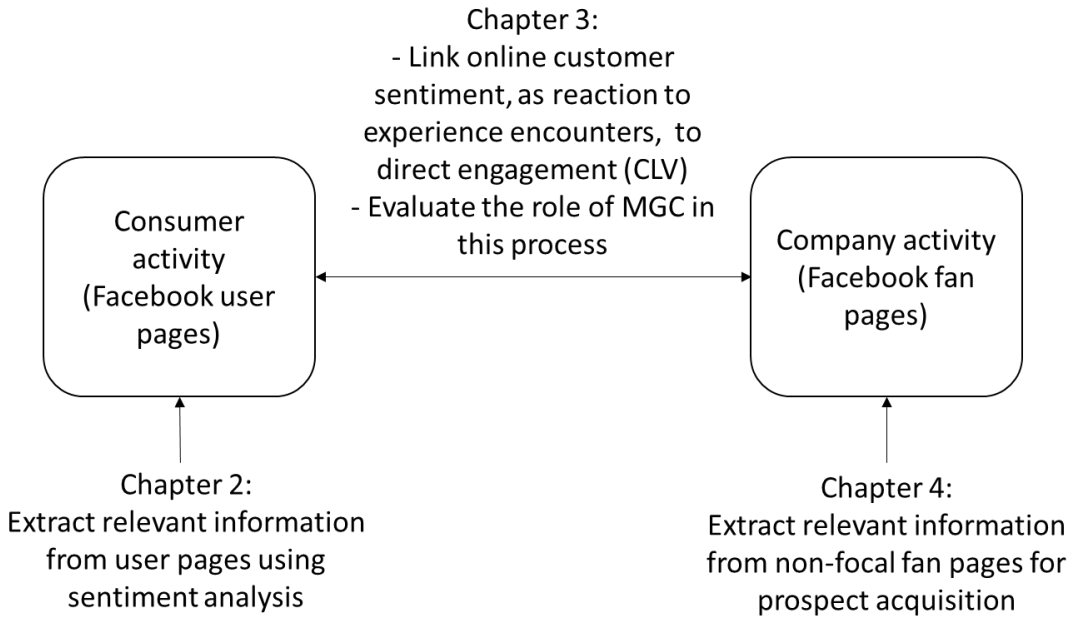


Figure 5.1: Graphical overview of this dissertation from a social media perspective

From the perspective of creating business value, the different chapters can also be seen as new (or extensions of existing) analytical approaches to CRM and to help evaluation the customer journey (cfr. Figure 1.2). Each of the chapters offers new approaches and insights that can complement existing research and applications, with a focus on data-driven marketing analytics.

2. Recapitulation of findings

2.1. Chapter 2

Electronic word-of-mouth (eWOM) has become widespread with the advent of social media. This offers opportunities for companies to monitor, assess, and use what is being told about them. Moreover, it has been shown that this online chatter results in increased sales, allows to monitor brand image and can be used in various other, non-marketing related topics. In this respect, online valence or sentiment prediction has become one of the main tools to evaluate eWOM. In chapter 2, we started from a traditional sentiment analysis which takes into account only textual characteristics (e.g., the text of a Facebook post). We proposed to enhance this model with leading and lagging information. Leading information is available before the text is posted on Facebook and includes user characteristics, but also previous user posts and their sentiment. Moreover, it allows to include deviations of the focal post from previous posts. Lagging information includes information that becomes available after the post has been published (e.g., Facebook likes and comments). The results show that adding leading information to the model substantially enhances model performance. Thus, previous post information and general personal characteristics can help to predict valence, even in real-time. In a last model, we added lagging variables to the model with textual characteristics and leading variables. Again, we see a substantial increase in model performance. It turns out that deviations from 'normal' posting behavior as well as comments and likes substantially increase our models' performance. We also see that the traditional textual information, leading and lagging information are all complementary and add to model performance in the most complete model. These results have high practical and academic value, since valence is commonly used in many fields.

2.2. Chapter 3

Existing research linking online customer content (eWOM or UGC) to company performance outcomes such as sales, tend to examine UGC and/or MGC over a particular period of time without aligning to a particular customer experience encounter. In chapter 3, we study UGC, in the form of online sentiment, related to actual customer experiences. Moreover, because we study actual experiences, we can assess the moderating role of (online) marketer generated content (MGC) on the link between the experiences' objective performance measures and the subsequent customer sentiment in SM. We further link individual customer sentiment to direct engagement (also known as customer lifetime value (CLV)), in combination with several

control variables linked to customer-firm interaction data. We compile a unique dataset in order to study the proposed model, comprising of SM data with several forms of UGC and MGC, objective performance measures for the customers' experiences, transactional data and marketing data. Using a two-phase model, we first show that MGC can effectively moderate the impact of the actual experience encounter on the displayed customer sentiment, and that MGC is particularly useful for more negative or neutral encounters. Next, the results show that customer sentiment has a positive and significant effect on direct engagement, even when controlling for transactional variables, and that this effect is relatively larger for purchase probability compared to purchase amount. Thus, MGC indirectly influences direct customer engagement through customer sentiment. Finally, we found that page likes on Facebook, arguably the most used metric on Facebook, is not significant for modeling customer engagement. With this paper, we are the first to link actual experiences, MGC, customer sentiment and direct engagement, thereby contributing to the growing literature streams of customer engagement and customer engagement management.

2.3. Chapter 4

The presence of companies on social media, e.g. a Twitter account or Facebook fan page, is not only a tool for these companies to interact with customers, it also reveals a lot of information about these companies. This information can then be used by other companies in their acquisition process. Despite the general interest in social media, also from business-to-business (B2B) organizations, only few have analyzed the impact of social media in the (B2B) sales process. Therefore, in chapter 4 we discussed the inclusion of Facebook page data of prospects into a customer acquisition model. More specifically, we devise a customer acquisition decision support system that includes the Facebook pages of prospects of Coca-Cola Refreshments Inc. (CCR), and compare the value of these social media data to commercially purchased prospect data and prospects' website data. Our system was subsequently used by CCR to generate calling lists of beverage serving outlets, ranked by their likelihood of becoming a customer. In this fourth chapter we report the results, in terms of prospect-to-customer conversion, of this real-life experiment encompassing nearly 9,000 prospects. The results show that social media data add value to predict prospect-to-customer conversion, over commercial and website data. Moreover, Facebook turns out to be the most informative data source to qualify prospects and is complementary with the other data sources. Finally, we argue that there can be a substantial monetary impact of using Facebook in an acquisition campaign in the proposed (quantitative) way.

3. Theoretical and managerial implications

The three studies in this dissertation offer several contributions, both to marketing theory and to marketing practice.

3.1. Theoretical contributions

From a theoretical perspective, we have studied the relatively under-researched area of social media marketing, and contributed to several aspects of this domain in the different chapters. In chapter 2, we introduced the notions of leading and lagging information for sentiment prediction models, which are promising paths to optimize sentiment prediction, next to research focusing on text elements. Relating to these variables, we laid out the fundamentals for more in-depth research into the formation of online sentiment, its antecedents and consequences. Although we do not formally test the proposed model, and especially the proposed middle-layer of unobserved concepts, we show the value of the observable characteristics in providing more accurate predictions of user sentiment. One option for future research would be to disentangle the effects and relationships of the unobserved concepts.

Chapter 3 makes significant contributions to the marketing literature, in several ways. First, we argue that online created content (UGC and MGC) can be linked to identifiable, actual customer experience encounters, instead of aggregating these measures over a particular period of time. This has important implications for our understanding of customer sentiment. We can link objective performance characteristics of the identified customer experience encounters to customer sentiment, and we can investigate the moderating role of MGC on the link between the experience encounter and customer sentiment. Second, we further link customer sentiment to direct engagement (CLV), thereby establishing the link between the experience encounters, MGC, customer sentiment and direct engagement in one model. We thus contribute to the literature on customer engagement (Pansari and Kumar, 2017) by demonstrating potential firm influences beyond more traditional marketing activities aimed at creating awareness. By doing so, we might link the theories of customer engagement (Pansari and Kumar, 2017) and customer engagement marketing (as conceptualized by Harmeling et al. (2017)). Whereas these latter authors focus on the direct influence of firm communications, our results support its moderating impact based on actual brand experiences. Third, we argue to include different measures of UGC and MGC in one comprehensive model, with control variables, in order to understand the influence of SM content on direct engagement, while previous literature has focused on individual measures. This allows researchers to better identify the real value of these social

media measures in relation to direct engagement. Finally, to the best of our knowledge we were among the first to introduce social media network metrics into direct engagement models, in addition to the other relevant social media variables. While previous research has focused mainly on networks via e-mailing or calling behavior (e.g., Nitzan and Libai, 2011), or has used social networks to set up viral marketing campaigns (Kumar et al., 2013), we show that social network information obtained via (online) social media also offer additional insights for modeling direct engagement with the firm. Thus, in spite of evidence stating that social media networks cannot readily be compared with offline networks because of the potentially large number of unrelated 'friends' (Dunbar, 2016), our research shows that the social media network is useful for modeling direct customer engagement.

Chapter 4 addresses the call for more (social media) marketing analytics research in B2B (Lilien, 2016). We are the first to quantitatively analyze the use of social media in the B2B acquisition process, instead of taking a qualitative approach. Moreover, from a modeling perspective, we have demonstrated that the acquisition model development is iterative in nature, and that it can benefit from including updated information into the model. With this research, we hope to spur academic interest in B2B applications in social media, since this is still an major untapped research topic.

3.2. Managerial contributions

From a managerial perspective, we have demonstrated in chapter 2 the ability to better predict customer valence related to Facebook posts. Since valence has been shown to be related to sales, it is important to correctly measure valence. Specifically in marketing, customer sentiment or satisfaction about a brand can be deduced from social media (e.g., Go et al., 2009; Schweidel and Moe, 2014; Tirunillai and Tellis, 2014). Making customer sentiment predictions more accurate also increases the applicability of these methods in comparison to previous methods (e.g., satisfaction surveys).

Chapter 3 offers insights for social media managers by investigating the role of MGC on social media. Our results imply that MGC can be effective to change customer sentiment, and ultimately customer engagement, but that its effectiveness is limited and dependent on the objective performance related to actual customer experience encounters. Positive customer experience encounters do not benefit as much from changes in MGC behavior as do more negative and neutral encounters. This is not surprising, since, within a service context, these latter encounters can be seen as service failures, and previous literature has already identified

that company-initiated recovery actions, such as MGC, can help to obtain service recovery (Smith et al., 1999). Moreover, the interactive nature of social media may further help to lead to positive service recoveries (Dong et al., 2008). Finally, we have shown that marketers' interest should go beyond merely measuring and influencing 'likes' on social media, to include (at least) customer sentiment.

Chapter 4 offers direction to B2B marketing managers in how social media can be used in a quantitative way. While we acknowledge that these models may be adapted to specific environments, we delineate a standard procedure to perform acquisition modeling, we show that social media is a valuable source of information in the context of prospect to customer conversion, and that this approach can be highly profitable.

4. Future outlook

Throughout this dissertation we have illustrated the potential of social media to create business value, and touched upon several interesting further research opportunities building on the presented research, such as the development of a theoretical framework for online sentiment creation, a deeper understanding of the role of MGC across different industries and applications, and more research on the use of social media in B2B-settings, both from a theoretical and marketing analytics point of view.

However, many more interesting questions regarding social media (value) remain unanswered to date. For instance, how consistent are the results over different industry types? How consistent are these results over different firm sizes? Which social media platform is most influential for which type of company? What about relatively newer social media such as Instagram, Pinterest and Snapchat and their influence? Which of the social media engagement actions of customers is most important for companies? Next to social media marketing through the social media pages of a company, other forms of social media marketing research continue to be important. Some of these streams (e.g. viral campaigns, influencer modeling) are already heavily researched (Aral and Walker, 2011; Berger and Milkman, 2012; Hinz et al., 2011; van der Lans et al., 2010), while other streams such as social media advertising received only little academic attention (Naylor et al., 2012; Tucker, 2014) and would benefit from more extensive research in order to understand how social media advertising works, to what extent it can increase meaningful firm outcomes and what may be necessary requirements for it in order to be effective.

All social media efforts can be seen as extra touchpoints with the company. These touchpoints become increasingly more difficult to control by the company, as social media are mainly driven by customers. However, social media also offer the opportunity to collect and measure many of these touchpoints. Combining both offline and online information (social media data, website data, internet-of-things related data) allows marketers to build more comprehensive models, and to better assess the relative value of each of these touchpoints. Synergies, spillover and crossover effects are likely to occur across different media and device types, and probably the type of media used might depend on the communication goal (i.e., convey a message or advertisement to a wide audience vs interaction with some customers). These insights could subsequently be used to get more complete insights in communication-mix elements, taking into account the value of touchpoints of the specific media and their specific roles. Thus, many research questions with high practical relevance are still on the table and provide promising avenues for future research (see for instance Wedel and Kannan (2016) for an overview of different research streams in Marketing Analytics).

However, social media also suffer from several potential pitfalls for future research. First, it becomes more and more difficult for companies to obtain social media data. Facebook, for instance, has already strongly tightened its API download policies. This makes it more difficult for both researchers and companies to obtain relevant social media. For instance, the data for the first study can still be collected, if a useful application is developed that uses the posts. A replication of the data for the second study is only partially feasible, since network data are not available anymore, and names of the comments cannot be retrieved anymore by the API. Chapter three data (fan page data) are still feasible to collect, since these are open data. Also social media data from Instagram, Pinterest and Snapchat are not easy to collect, which means companies have to resort to their own collection and statistics (which are often not very detailed). Second, and related to the first point, privacy issues become more and more prevalent (Baesens et al., 2016). Customers are more cautious to share new information, and at the same time social media tools are more restrictive to share information. Moreover, governments are putting in place strict privacy legislations that prescribe and limit the use of personal and detailed information. In the European Union for example, the right to be forgotten will soon be in practice (Macaulay, 2017), and the recently introduced and much bespoken external regulation in the form of GDPR. As a consequence, future marketing-mix (or other types of) models should be designed to cope with privacy regulations limitations and be able to handle anonymized and minimized data (Wedel and Kannan, 2016). While this may limit the practical

implementation of the proposed models, the main insights that come from these studies already offer more in-depth understanding of the working mechanisms and importance of social media which is important given the enormous amount of money spent on social media nowadays.

Social media offer the potential to collect data on individuals, but not every customer is a social media user. Thus, there are limitations to the generalizability of the results found using social media. Put in another way, working with social media often lead to selection effects. This is even more present when using mobile application users, who basically self-selected into a study (e.g., using a Facebook application as in Chapter 3). In this case, we need to accordingly adjust the analysis, for instance with Propensity Score Matching or a Heckman selection model. However, the increasingly complex models cannot easily be adapted to include these corrections (at least the Heckman correction). For instance, a combination of panel data with a binary selection and outcome variable already proves to be a serious challenge that has only just been resolved (Semykina and Wooldridge, 2017). Therefore, it is important that these modeling issues will be further resolved to make full use of the social media data.

5. References

Aral, S., Walker, D., 2011. Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks. Management Science 57, 1623–1639.

Babić Rosario, A., Sotgiu, F., De Valck, K., Bijmolt, T.H.A., 2016. The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. Journal of Marketing Research 53, 297–318.

Baesens, B., Bapna, R., Marsden, J., Vanthienen, J., Zhao, J., 2016. Transformational issues of big data and analytics in networked business. MIS Quarterly 40, 807–818.

Berger, J., Milkman, K.L., 2012. What Makes Online Content Viral? Journal of Marketing Research 49, 192–205.

Chen, Y., Xie, J., 2008. Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix. Management Science 54, 477–491.

Dong, B., Evans, K.R., Zou, S., 2008. The effects of customer participation in co-created service recovery. J. of the Acad. Mark. Sci. 36, 123–137.

Dunbar, R.I.M., 2016. Do online social media cut through the constraints that limit the size of offline social networks? Royal Society Open Science 3, 150292.

Go, A., Bhayani,R., Huang,L., 2009. Twitter sentiment classification using distant supervision, Technical report, CS224N Project Report, Stanford, 2009.

Harmeling, C.M., Moffett, J.W., Arnold, M.J., Carlson, B.D., 2017. Toward a theory of customer engagement marketing. J. of the Acad. Mark. Sci. 45, 312–335.

Hennig-Thurau, T., Malthouse, E.C., Friege, C., Gensler, S., Lobschat, L., Rangaswamy, A., Skiera, B., 2010. The Impact of New Media on Customer Relationships. Journal of Service Research 13, 311–330.

Hinz, O., Skiera, B., Barrot, C., Becker, J.U., 2011. Seeding Strategies for Viral Marketing: An Empirical Comparison. Journal of Marketing 75, 55–71.

Kumar, V., Bhaskaran, V., Mirchandani, R., Shah, M., 2013. Practice Prize Winner—Creating a Measurable Social Media Marketing Strategy: Increasing the Value and ROI of Intangibles and Tangibles for Hokey Pokey. Marketing Science 32, 194–212.

Lilien, G.L., 2016. The B2B Knowledge Gap. International Journal of Research in Marketing 33, 543–556.
Macaulay, T., 2017. What is the right to be forgotten and where did it come from? | Data | Techworld [WWW Document]. URL https://www.techworld.com/data/could-right-be-forgotten-put-people-back-in-control-of-their-data-3663849/ (accessed 1.8.18).
Naylor, R.W., Lamberton, C.P., West, P.M., 2012. Beyond the "Like" Button: The Impact of Mere Virtual Presence on Brand Evaluations and Purchase Intentions in Social Media Settings. Journal of Marketing 76, 105–120.
Nitzan, I., Libai, B., 2011. Social Effects on Customer Retention. Journal of Marketing 75, 24–38.
Pansari, A., Kumar, V., 2017. Customer engagement: the construct, antecedents, and consequences. Journal of the Academy of Marketing Science 45, 294–311.
Schweidel, D.A., Moe, W.W., 2014. Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. Journal of Marketing Research 51, 387–402.
Semykina, A., Wooldridge, J.M., 2017. Binary response panel data models with sample selection and self-selection. J Appl Econ. 2017, 1–19
Smith, A.K., Bolton, R.N., Wagner, J., 1999. A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery. Journal of Marketing Research 36, 356–372.
Tirunillai, S., Tellis, G.J., 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research 51, 463–479.
Tucker, C.E., 2014. Social Networks, Personalized Advertising, and Privacy Controls. Journal of Marketing Research 51, 546–562.
van der Lans, R., van Bruggen, G., Eliashberg, J., Wierenga, B., 2010. A Viral Branching Model for Predicting the Spread of Electronic Word of Mouth. Marketing Science 29, 348–365.
Wedel, M., Kannan, P. k., 2016. Marketing Analytics for Data-Rich Environments. Journal of Marketing 80, 97–121.
Zhang, Y., Moe, W.W., Schweidel, D.A., 2017. Modeling the role of message content and influencers in social media rebroadcasting. International Journal of Research in Marketing 34, 100–119.