The Incremental Value of Unstructured Data in Predicting Customer Churn

Evert de Haan and Elena Menichelli
The Incremental Value of Unstructured Data in Predicting Customer Churn

Evert de Haan\textsuperscript{a}
Elena Menichelli\textsuperscript{b}

August 1, 2019

\textsuperscript{a}Department of Marketing, Faculty of Economics and Business, University of Groningen, Nettelbosje 2, 9747AE Groningen, phone: +31 50 36 32802, email: evert.de.haan@rug.nl. (corresponding author)

\textsuperscript{b}Department of Finance and Business Management, University of Stavanger, Kjell Arholms gate 41, Postboks 8600 Forus, 4036 Stavanger, Norway, phone: +47-51832994, email: elena.menichelli@uis.no.

The authors thank the anonymous data provider for granting access to the data. They also thank the participants of the 47th EMAC Conference in Groningen, the 8th Theory + Practice in Marketing Conference in Los Angeles, and the 40th ISMS Marketing Science Conference in Philadelphia for their valuable feedback.
INTRODUCTION

While companies’ strategies have traditionally focused on selling products and services with limited knowledge about their customers, today’s firms have become more “customer centered” (Abdolvand, Albadvi, and Aghdasi 2015; Tosti and Herbst 2014). Understanding consumers’ expectations and meeting or even exceeding these expectations to enhance customer satisfaction can boost a firm’s survival rate (Lemon and Verhoef 2016) and help improve the position of the marketing department within the firm (Verhoef and Leeflang 2009). Collecting rich data on and gaining insights into customers are thus imperative for firms, as superior firm performance is an outcome of satisfying customers more effectively than competitors (Gupta and Zeithaml 2006; Osborne and Ballantyne 2012).

Customer relationship management (CRM) typically uses market(ing) information and customer data, such as demographics and product or service usage information (for a review, see Kumar and Reinartz 2016), to predict future customer behavior, including the likelihood to respond to an offering and the likelihood to churn (Lemmens and Croux 2006; Neslin et al. 2006). These customer-level data are usually collected during customers’ lifetime, from the moment of signing until the moment of churn. Then, firms use these data, which are stored in the customer database, for marketing decision making, a process called “database marketing” (Blattberg, Kim, and Neslin 2008).

In addition to database marketing, literature has focused on the usage of customer feedback metrics (CFMs), including customer satisfaction and the Net Promoter Score (NPS; Reichheld 2003). These data, which typically come from surveys, are good predictors of future customer behavior, including customer retention (De Haan, Verhoef, and Wiesel 2015). This type of information provides incremental predictive value beyond traditional data from the customer database.
Moreover, beyond the structured data typically found in customer databases and CFMs collected through surveys, a large amount of unstructured data about a firm are available from customers. Examples of such unstructured data are written and verbal opinions, often collected through open-ended survey questions, (email) correspondence between the firm and the customer, and online comments found on social media and review sites. According to a report from International Data Corporation, in 2013 the overall data volume in the world was 4.4 zettabyte (4.4 trillion GB), with this size likely to double every two years (Turner et al. 2014). Most of these data are unstructured (e.g., text, audio, images, videos), with only .5% being analyzed (Burn-Murdoch 2012). A reason for this low analysis rate is that most analytical methods are developed for structured data only and cannot handle unstructured data (for a discussion of methods, see Blattberg, Kim, and Neslin 2008).

Unstructured data can however contain valuable information for churn prediction which isn’t available in structured data, such as a personality traits and mental state (Netzer, Lemaire, and Herzenstein 2017), or information about who is at risk to churn (e.g. by looking at the emotions or sentiment of a text) and why someone is at risk to churn (e.g. by examining the topic of a text), as indicated by Ascarza et al. (2017). Thus, a question we aim to address in this research is how valuable unstructured data are for CRM, in terms of both making better predictions and determining the incremental monetary value of these data.

More specifically the aims of this article are threefold: (1) to investigate ways to extract insights from unstructured (written) data that may be relevant for firms; (2) to add the extracted information from these data to econometric models and dashboards, so that it can be systematically exploited for customer management purposes; and (3) to uncover the extent to which qualitative data can provide richer insights into the customer database and what the incremental (monetary) value of these data is. The Marketing Science Institute (2016, 2018) specifies the increasing need to include and handle new data sources in its research priorities.
from both 2016–2018 (“New data, new methods, and new skills—how to bring it all
together?”) and 2018–2020 (“What approaches exist to capture and analyze non-structured
data … to improve firm communications and customer experience?”). Our study also
addresses a direction for further research that Wedel and Kannan (2016) note: “How can firms
derive individual-level insights from big data … to give readings of customers’ intentions in
real time?” as well as Ascarza et al. (2018) “What are the best ways to leverage unstructured
data (…) to enhance retention management?”

In our empirical application, we analyze data from 50,000 customers from a European
telecom provider. We show that the simultaneous handling of different types of information
improves churn prediction. We also demonstrate that all three considered data sources—
namely, data from the customer database, the NPS, and data extracted from written customer
feedback—present unique information that assists in predicting customer retention. The data
extracted from written customer feedback have an incremental monetary value of $315,000
annually for our data provider. In addition, these data can help the provider to better
understand the reasons some customers churn and some stay, which can improve customer
targeting and thus enhance customer service and customer experience. We furthermore show
how these data can be included in customer dashboards and also provide potential insights for
the entire customer base, beyond only those who have provided written customer feedback.

The rest of this article proceeds as follows: We begin with a review of the literature, in
which we focus on typical customer database data, CFM data, unstructured data, and the
potential benefits of combining these three data sources. After this, we discuss our empirical
setting by describing the data, the methods we use to extract information from written
customer feedback, and our methodology to compare and combine the three data sources.
Next, we examine the results and how written customer feedback can be used in customer
dashboards. Finally, we conclude with a discussion of the limitations and directions for future research.

LITERATURE REVIEW

The marketing literature is replete with studies that try to predict future customer behavior using data from customer databases. Another stream of literature focuses on predicting firm performance and customer behavior using CFMs collected through surveys. Finally, a more recent field in marketing literature centers on extracting information from unstructured data (e.g., textual feedback from customers). We review these three literature streams in more detail and then discuss how combining the different information sources can be valuable for firms in making better predictions of future customer behavior.

Customer Database Management

When trying to forecast future customer behavior, firms and researchers have usually used data from customer databases (Blattberg, Kim, and Neslin 2008). In particular, to predict the future value of customers, in terms of net profit attributed to future customer relationships, they have used measures such as recency, frequency, and monetary value (RFM) (e.g., Fader, Hardie, and Lee 2005; Levin and Zahavi 2001). More recently, research has assessed customers’ future revenue and growth rates using past information about customers to predict the customer lifetime value (CLV) (e.g., Niraj, Gupta, and Narasimhan 2001; Venkatesan and Kumar 2004).

Another important forecasting of customer behavior is customer retention, which involves the activities and actions companies take to reduce the number of customer defections, or churns. Gupta, Lehmann, and Stuart (2004) show that customer retention is the most important component of customer equity, which in turn has an almost one-to-one relationship to a firm’s market value. Effectively predicting which customers will leave and
which ones will stay is therefore an important topic in marketing. For example, in their churn competition, Neslin et al. (2006) use different methodologies to forecast which method will best predict customer churn at a telecom provider. The best-performing methods to predict churn in this competition were logistic regression models and decision trees (Lemmens and Croux 2006; Neslin et al. 2006). The individual variables considered to predict churn are typical customer database variables, such as customer behavioral data (e.g., usage, revenue), company interaction data (e.g., calls to the customer service center), and customer household demographics (e.g., age, income, geographic location, home ownership) (Neslin et al. 2006). Studies use similar data when predicting customer churn. For example, Risselada, Verhoef, and Bijmolt (2010) and Holtrop et al. (2017) use data from customer databases to show the staying power of churn models (i.e., how well the models using these data predict churn and how long these predictions stay accurate). Again, all these methods perform well in predicting churn with different fit criteria and can predict reasonably well ahead. An important question remaining, however, is whether data from customer databases are sufficient or if other data sources should be added to analyses to obtain even better predictions.

CFMs

A wide range of studies have investigated how CFMs, also called “mindset metrics,” help explain and predict firm performance and customer behavior (for an overview, see Gupta and Zeithaml 2006). One of the most researched metrics is customer satisfaction, which is positively related to both firm performance and customer behavior. For example, Morgan and Rego (2006) find that customer satisfaction is positively related to Tobin’s q, cash flow, shareholder returns, sales growth, gross margin, and market share. All these factors have a positive impact on shareholder value (Gruca and Rego 2005), making customer satisfaction a good metric for portfolio management (Aksoy et al. 2008; Anderson, Fornell, and Mazvancheryl 2004). Research has also found that customer satisfaction is positively related
to customer behavior, with implications for (future) expenditures (Ittner and Larcker 1998),
cross-buying and share of wallet (Loveman 1998), repurchasing (Mittal and Kamakura 2001),
and retention (De Haan, Verhoef, and Wiesel 2015; Ittner and Larcker 1998).

In addition to customer satisfaction, another relevant CFM is the NPS (Reichheld 2003),
which is one of the most popular metrics applied by practitioners (Bain & Company 2016). Although the NPS received criticisms in the literature after its introduction (e.g.,
Keiningham, Cooil, Aksoy, et al. 2007; Keiningham, Cooil, Andreassen et al. 2007; Morgan
and Rego 2006), more recent studies measuring the NPS have found that this metric is as
equally valuable as customer satisfaction in predicting firm performance (Van Doorn,
Leefflang, and Tijs 2012) and customer retention (De Haan, Verhoef, and Wiesel 2015).

A drawback of CFMs is that while the variables used in traditional database marketing
are often available for all customers, CFMs are only available for a subset of customers—
namely, those who provided feedback through a survey or during a service encounter. These
CFMs tend to be less widely available but nonetheless provide incremental predictive value.
De Haan, Verhoef, and Wiesel (2015) show that CFMs help predict customer retention more
precisely than when using only information on relationship length and customer
demographics. A reason is that the information from customer databases is mainly on
customer characteristics and historic usage and interactions (Blattberg, Kim, and Neslin 2008;
Neslin et al. 2006), while CFMs exploit customer attitudes and perceptions. Combining these
types of data can therefore lead to better predictions (De Haan, Verhoef, and Wiesel 2015).

Unstructured Data

In the past 20 years, data has increased in large scale in various fields. Compared with
traditional data, big data typically includes a massive amount of unstructured data. On the one
hand, big data has an enormous potential for gaining in-depth insights; on the other hand, it
also incurs new challenges, such as how to organize and manage such data sets effectively (Chen, Mao, and Liu 2014).

Unstructured data in marketing include, for example, customers’ online product reviews, blogs, comments on social media, verbal or written interactions with firms, and so on. This information is qualitative in nature and cannot directly be used in quantitative analysis, which dominates in the field of customer management. Instead, the data need to be preprocessed before being combined with other quantitative information. For textual data, the preprocessing phase includes (1) feature selection, to highlight features that can represent a piece of text through, for example, the “bag-of-words” technique (Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos 2015; Zhang, Jin, and Zhou 2010); (2) dimensionality reduction, to keep most of the variability in few variables, as increasing the number of features decreases the efficiency of the clustering or prediction model (Crain et al. 2012; Miner, Elder, and Hill 2012); and (3) feature representation, to represent each feature by a quantitative value by, for example, occurrence frequency (Dumais et al. 1998; Jiang, Li, and Huang 2016).

The next step after structuring the qualitative information is to extrapolate relevant insights from text, such as customer sentiments and opinions or main topics. A plethora of approaches for doing so is available (for a review, see Nikolenko, Koltcov, and Koltsova 2017; Serrano-Guerrero et al. 2015). Sentiment analysis, or opinion mining, involves detecting the emotional sentiment preserved in text by means of a semantic study (Nassirtoussi et al 2014). This is useful, for example, to gauge the market reception of a new product, to evaluate overall customer feedback, or to estimate the popularity of a brand (Ghiassi, Skinner, and Zimbra 2013; Mostafa 2013). The body of research focusing on sentiment analysis (e.g., Balahur et al. 2009; Cambria et al. 2013; Chen and Zimbra 2010) is mainly based on identifying positive and negative words in a text and classifying the text’s
emotional stance as positive or negative. For example, Maks and Vossen (2012) present a lexicon model for deep opinion mining, and Desmet and Hoste (2013) examine emotion detection in suicide notes.

Topic modeling refers to quantitative techniques that provide users with an overview of themes being discussed in documents. Hierarchical probabilistic models can help find patterns of words in document collections. Initially, topic models served as an information retrieval tool, intended to make browsing large collections of documents easier (Salton, Wong, and Yang 1975); today, topic modeling methodology is a vibrant research subject. The two most common model types used in applied work are latent semantic analysis (Dumais et al. 1998) and latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003).

In the marketing literature, text-mining methods have become more applied in the last few years. For example, Tirunillai and Tellis (2012) investigate how the sentiment of online chatter relates to abnormal stock return and trade volume, Salehan and Kim (2016) examine the effect of the sentiment of a review on readership and perceived helpfulness of online reviews, Zhang, Moe, and Schweidel (2017) show how the content of a message (tweet) encourages people to rebroadcast (i.e., retweet) the message, and Netzer, Lemaire and Herzenstein (2017) show that including textual information help improve models in predicting loan default. The study that comes closest to our study is however by Coussement and Van den Poel (2009), who include both CRM and textual information to predict churn, but the focus of their study is on the best model to predict churn, while we are focused on the incremental value of different data sets.

Table 1 provides an (non-exhaustive) overview on how research has used (combinations of) the three different data sources in CRM settings. As the table shows, research often uses typical customer database variables to predict customer-level (behavioral) outcome variables, in some cases also combined with data collected through surveys. The
research stream of unstructured data is however a separate field of research that struggles to be incorporated in CRM-related research, with a few notable exceptions. Our study is to the best of our knowledge the first to incorporate and compare all three data sources simultaneously, and the first to investigate the incremental (predictive and monetary) value of each data source. We furthermore investigate how the impact of the different variables on churn (or retention), our dependent variable of interest, changes over time. This investigation sheds light on the strength of the impact of each variable and on how far ahead it can make accurate and significant predictions.

DATA

A large European telecom provider provided us with a data set containing information from 50,000 customers who contacted the customer service center between January 2013 and November 2014 and replied to a short message service (SMS) survey. These 50,000 customers are a random, balanced sample consisting out of ~50% churners. Each time a customer contacts customer service (e.g., about a question, complaint, or request), the provider sends him or her an SMS with the NPS questions (Reichheld 2003). In addition, customers are asked to explain their reasons for their evaluations by writing a short text. The text is in the local (non-English) language. In our study, the NPS is our CFM data of interest, and the written customer feedback is the unstructured text data.

In addition to this information, we gathered other customer information from the customer database, including the customer–telecom provider relationship length (up to the time of the service call) and customer demographics. Details about the service call were also available, such as the time the customer waited in the queue before the call was answered by a service employee and the length of the actual conversation. In addition, the database details
whether customers churned before February 16, 2015, and, if so, on which day they churned; these two pieces of information served as the dependent variables in our study.

The database also included customers who contacted the customer service center after they churned (e.g., customers who still had remaining issues that needed to be addressed). However, because we wanted to predict future churn behavior, we excluded these customers (n = 5,640) from our data set. In addition, we dropped customers who stated that they were younger than 18 years of age (n = 93), because these customers need their parents’ permission for a telecom subscription and are not the decision makers themselves. Finally, we deleted customers who, after we cleaned the textual response, did not have any text remaining. These are mainly textual responses that include only a sign (e.g., a series of dots) or a number (e.g., the provided NPS), which are not useful for our analyses. This resulted in the exclusion of 109 observations. Overall, we dropped 5,808 of the 50,000 observations, which left us with 44,192 customer observations.

Table 2 lists the variables and their descriptive statistics. The CRM variables we have chosen are similar to the ones included in other studies that aim at predicting churn (e.g. Neslin et al. 2006) and are recommended by textbooks (e.g. Blattberg, Kim, and Neslin 2008), i.e. variables on demographics, services used, service history, service interaction, as well as moment of interaction. With this we try to estimate a good baseline model suggested by literature and also in line with how our data provider predicts churn. As indicated in Table 2, our sample is balanced with regard to churn and gender. Regarding the length of the relationship with the firm, the data set is heterogeneous in terms of new customers and customers who have been with the provider for a long time. The same holds true for customer age and NPS evaluations, as well as the services customers use and the time of the service call (i.e. the time of day, day of the week, month and year in which the service call took place). For the number of minutes waiting in the queue and the number of minutes the conversation
lasted, we find high skewness; most customers waited only for a short while in the queue (e.g., 82.50% of customers waited for less than one minute), while in the most extreme case, customers waited for more than 49 minutes. Similarly, most conversations were rather short (e.g., 86.84% of the conversations were less than ten minutes), while the longest conversation was just over 165 minutes. Because of this skewness, we log-transformed the variables years_at_firm, minutes_in_queue, and minutes_conversation in our analyses. Finally we also see that most customers (95.0%) have had multiple calls with the customer service center and 16.3% have been called back by the firm after the initial service call.

Table 3 provides the correlations between the variables; the first column indicates that all variables are significantly correlated with churn. The years_at_firm variable (i.e., relationship length) has the strongest correlation with churn: the longer the customer has been with the firm, the lower is the churn probability. The NPS variable is also negatively correlated with churn. Both correlations, as well as their magnitudes, are in line with the findings of De Haan, Verhoef, and Wiesel (2015), who investigate the relationship of NPS and other CFMs with customer retention. In addition, we find that women, older people, customers who had a longer conversation with the customer service center, as well as those who have been called back are less likely to churn. Surprisingly, minutes_in_queue is negatively correlated with churn. Given that waiting time is also correlated with some of the other variables, this might explain this counterintuitive correlation, which we investigate subsequently.

**TEXT ANALYSES**

A challenge when using unstructured data (e.g., written feedback) is that such data cannot be included directly in traditional statistical methods. As discussed in our literature review, however, text can contain rich information. In our empirical setting, the written
feedback can (1) indicate something about the willingness of the customer to provide detailed feedback (e.g., the length of the written response); (2) shed greater light on a specific NPS and its actual meaning, in terms of positive or negative experiences (e.g., some people might give an NPS of 7 to express satisfaction with the provider, while for others, a 7 might be an indicator that something went wrong; see Van Doorn, Leeﬂang, and Tijs 2013); and (3) indicate why the customer provided a speciﬁc form of feedback (e.g., the reason for the call or the NPS). In the next subsections, we delineate how we handled our textual data and the information we extracted from them to predict customer retention.

Text Data Structuring

As indicated previously, all written feedbacks are in a non-English language. This has a few consequences especially for the sentiment analyses, because we must either use a sentiment dictionary in the corresponding language or translate the text into English to be able to use an English sentiment dictionary. Given that no sentiment dictionary is available in the original language in the package we use (i.e., the R-package “qdap”), we tested alternative options to determine which worked best. First, we used Google Translate to translate the existing English sentiment dictionary to the local language and then performed sentiment analysis using this newly created sentiment dictionary. As a second option, we translated the original customer texts into English and used the original English sentiment dictionary to calculate the sentiment scores. The sentiment scores in both cases were comparable and highly correlated with each other (r = .750). Given that the sentiment scores from the procedure in which we translated the customer texts are more highly correlated with churn (r = −.108) than the procedure in which we translated the sentiment dictionary (r = −.095), we used the former in our analyses.

To correctly handle the textual data for sentiment and topic analyses, we needed to process and clean the textual data. For this purpose, we undertook seven steps:
After translating and cleaning the text, we conducted the sentiment and topic analyses.

**Sentiment Analysis**

In addition to the structure and specific words used in a text, the sentiment of the text can help predict churn. For example, two sentences might contain the same (combination of) words (e.g., “satisfied” and “service employee”), but one might be framed positively (“I was satisfied with the service employee”) and the other negatively (“I was not satisfied with the service employee”). For predicting future customer behavior, the sentiment in which something is said is likely more important than the individual words being said (i.e., whether what is said has a positive or negative slant).

Different approaches can be used to measure the sentiment of the text documents. One possible method employs hand coding (e.g., one or multiple people read the text and rate whether it is positive, negative, mixed, or neutral). A downside of this approach is its subjective nature, especially when one person codes the text. Another downside is that in practice, a firm might have an enormous amount of textual data (e.g., from social media or
customer surveys), making this process expensive, impractical, and, most of the time, not in real time.

Another method, which is based on automated text-mining tools, determines the degree to which a text is positive or negative and then applies a so-called sentiment or polarity score to the text. The advantage of this procedure is that it is inexpensive, objective (the same comments will receive exactly the same score every time), and can be run in real time. Furthermore, it increases replicability and comparability across studies and situations (e.g., customer groups, regions, product types, industries), which can be both scientifically and managerially desirable.

A disadvantage of using algorithms is that it is more difficult to understand complicated language nuances such as sarcasm or double negatives, resulting in inaccurate polarity scores. Especially in the case of very rough sentiment analysis techniques, which investigate the number of positive and negative words contained in a text to determine the sentiment (e.g. as done by Coussement and Van den Poel (2009) and suggested by Ascarza et al. (2019)), this creates several limitations. More advanced sentiment analysis techniques, however, not only examine (the number of) positive and negative words used in a text but also assess the context in which these words are used. Again, in our study, we use the R-package qdap (v. 2.3.0), which standardly includes both polarity score calculations and the context of a text (for more details on this package, see Rinker 2018). To calculate the polarity score for each sentence, qdap first uses the sentiment dictionary provided by Hu and Liu (2004) to tag polarized words (i.e., positive and negative). For this study, to reduce subjectivity, increase replicability, and demonstrate the usability of this package, we use the standard settings. Examples of polarized words are “horrible” and “incredible.” From around each polarized word, we pull a context cluster ($x_i^T$) of words, which consists of (in the standard settings of qdap) four words before and two words after the polarized word. The words in the context
cluster can be tagged as neutral ($x^0_i$, not affecting the sentiment of the polarized word), negator ($x^N_i$, inversing the sentiment; e.g., “not” or “don’t” as in “I don’t like ...”), amplifier ($x^a_i$, increasing the sentiment; e.g., “seriously” as in “I seriously like ...” or “I seriously don’t like ...”), or deamplifier ($x^d_i$, decreasing the sentiment; e.g., “barely” as in “I barely like”). These valance shifters can thus inverse (negator), strengthen (amplifier), or weaken (deamplifier) the polarized word (Rinker 2018).

To calculate the polarity score, after identifying each polarized word and its context cluster(s), we check the text for negation (or double, triple, and so on, negation) in the context clusters using Equation 1:

$$w_{neg} = (\sum x^N_i) \mod 2,$$

where $\sum x^N_i$ is the number of negator words. This simply means that $w_{neg}$ has a value of 1 when there are an uneven number of negators in the context cluster and 0 when there is an even number of negators. After this, we calculate the degree to which the context cluster is deamplifying, which has a minimum value of $-1$:

$$x^d_i' = \sum (-w_{neg} \cdot x^a_i + x^d_i),$$

$$x^d_i = \max(x^d_i', -1),$$

where $x^a_i$ and $x^d_i$ are the number of amplifiers and deamplifiers in the context cluster, respectively. Equation 2 indicates that if there is negation in the context cluster, as defined by Equation 1, an amplifier turns into a deamplifier. In other words, the combination of a negator and an amplifier (e.g., “not very helpful”) has the same meaning as a deamplifier (e.g., “barely helpful”).

Subsequently, we calculate the degree to which the context cluster is amplifying:

$$x^a_i = \sum (-w_{neg} - 1) \cdot x^a_i).$$

This indicates that when there is negation in the context cluster (as defined by Equation 1) together with an amplifier, the amplifier does not have an amplifying effect (i.e., is set to
zero). An amplifying effect only emerges when there is no negator (as defined by Equation 1) in the respective context cluster.

After this, we can calculate the polarity score per context cluster and sum up each using Equation 5:

$$x_i^T = \sum((1 + c(x_i^A - x_i^D)) \cdot w(-1)^{\sum x_i^N}),$$

where $c$ is the weight of the (de)amplifying effect (which in qdap is set at .8) and $w$ is the value of the polarized word (which in the standard setting is set at +1 for positive words and –1 for negative words). In practice, analysts could fine-tune this to a greater extent (e.g., “great” can have a higher value $w$ than “good,” “horrible” can have a more negative value than “bad”).

We then calculate the final polarity score of a text using Equation 6, which corrects for the length of the text measured in number of words $n$. This ensures that a long text with a few positive (or negative) context clusters has a lower overall polarity score (in absolute value) than a short text with an equal number of positive (or negative) context clusters (for full details, see Rinker 2018):

$$\delta = \frac{x_T^r}{\sqrt{n}}.$$

For sake of clarity, we provide a few examples of the polarity scores in Table 4. As the table shows, the word “happy” is as positive as the word “unhappy” is negative, resulting in an inverse polarity score for the second sentence. “Not happy” is equivalent to the positive word “happy” with the negator “not” in front of it, resulting in an inverse score as well (see the first three sentences in Table 4, which have the same number of words). The phrase “barely happy” contains a deamplifier, which lowers the positive impact of the polarized word “happy”. The same is shown for “barely unhappy”, which lowers the negative impact of the polarized word “unhappy”. Finally, “very (un)happy” contains an amplifier, which makes the polarity score higher than when someone is just “(un)happy.” To keep the number of words
constant in Table 4, we removed the neutral word “provided” in sentences three to six. Keeping this word in the sentence, as in sentences seven through ten, increases the number of words and thus affects the polarity score, given that a larger part of the text is neutral. We calculate all the polarity scores in Table 4 using the “polarity” function in the R-package qdap (v. 2.2.5).

As indicated previously, we used the standard setting and dictionaries by qdap to calculate the polarity score of all 44,192 written feedbacks. Then, we randomly selected 500 written feedbacks and asked two coders to hand-code each text as (more) positive, neutral, (more) negative, or mixed. Next, we compared the human coded texts with the polarity scores calculated by qdap, which can be positive, neutral/mixed (i.e., a polarity score of zero), or negative. The two matched in 76.6% of the cases. In only 4.8% of the cases were the texts coded in the opposite way (coded as negative by the human coders and but positive by qdap, and vice versa). Using the nonmatched cases, we updated the sentiment dictionaries, including an additional 48 positive words and 28 negative words to the original qdap dictionary. This improved the total hit rate to 80.6%, with only 3.8% being coded in the opposite way. For our analyses, we use this improved dictionary.

Figure 1 shows the histogram of the final polarity scores. The polarity score of the provided text is positive in 63.72% of the cases and negative in 17.46% of the cases; in 18.83% of the cases, the polarity score is zero (i.e., the text is neutral or mixed). The correlation of the polarity score with churn is –.103 (i.e., the more positive the sentiment of the text, the less likely the customer is to churn). If we use effect coding (+1 for positive, 0 for neutral/mixed, and –1 for negative), the correlation is –.108. Therefore, we use this effect coding of polarity for later analyses.
**Topic Analysis**

LDA is a Bayesian mixture model for discrete data that classifies a text into different topics (Blei, Ng, and Jordan 2003). To do this, we extracted all unique words in our text. The 44,192 written feedbacks, in the original language, contained 11,242 unique words, after we applied the cleaning procedure. Because some words only appear in one or a few texts, which makes LDA difficult and can result in many unique topics, we kept only words that appeared in at least 1% of all texts (i.e., in at least 442 of the 44,192 texts). This resulted in 127 unique words. Creating a data frame of $127 \times 44,192$, we applied Grün and Hornik’s (2011) three steps:

1. The term distribution $\beta$ is determined for each topic by $\beta \sim Dirichlet(\delta)$;
2. The proportions $\theta$ of the topic distribution for the text file are determined by $\theta \sim Dirichlet(\theta)$;
3. For each of the $N$ words $w_i$:
   
   a. Choose a topic $z_i \sim Multinomial(\theta)$;
   
   b. Choose a word $w_i$ from a multinomial probability distribution conditioned on the topic $z_i$; $p(w_i|z_i, \beta)$.

For this, we used the R-package “topicmodels” (v. 0.2-7). Because the number of topics needs to be predetermined for our modeling, we ran the LDA procedure for topics between 1 and 50; using the criteria that Cao et al. (2009) and Deveaud, SanJuan, and Bellot (2014) set, we found that the resulting optimal number is 6. Given that some written feedbacks also do not contain any of the 1% most-frequently-used words, we do not assign them to a specific topic (which is our base case). Table 5 shows all the topics, their size, the top five most frequently used words (translated into English), the polarity scores, and the churn rate.
Given the top five words, the first topic is call(ing) related; in addition, the related written feedbacks are quite mixed in terms of polarity, and this topic has a rather high churn rate (49.1%). In total, 16.6% of all written feedbacks belong to this topic. We also find that feedback about customer service, speed/coverage, and quick help (i.e., topics two, three, and four) are generally positive in terms of sentiment, and the churn rate is relatively lower.

**METHODOLOGY**

In our study, we aim to investigate the extent to which the three different data sources (i.e., the customer database data, the structured survey data, and the unstructured textual data), can predict churn. We make predictions beginning from the moment the customer contacted the provider’s service center until the moment of churn, which is right-censored at February 16, 2015 (i.e., we only observe churn up until this date). Because of the structure of our data with right censorship (churn that is not observed can still take place in the (distant) future) and since observed churn can be right after the service call but can also take a longer period of time, we use a Cox proportional hazards model (Kleinbaum and Klein 2012). We calculate the basic structure of this model as follows:

\[
h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i},
\]

where \( h(t, X) \) is the hazard probability on day \( t \) given the independent variables \( X \), \( h_0(t) \) is the baseline hazard on day \( t \), and \( X_i \) are the independent variables. Because some of the variables might have a different impact over time, we allow \( \beta_{it} \) to be time dependent (i.e., become more or less strongly related to the hazard probability over time after the service call).

We estimate the eight versions of this model as follows:

1. A baseline model with no independent variables (i.e., assuming everyone has the same hazard function \( h_0(t) \));
2. Three models containing one of the three data sources as the independent variable;
3. Three different models containing a combination of two of the data sources as independent variables; and

4. One model containing all three data sources as independent variables.

We compare the eight models using the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Similar to De Haan, Verhoef, and Wiesel (2015), we follow Wagenmakers and Farrell’s (2004) procedure and calculate the Akaike weights to determine which of the models—that is, which (combination of) data source(s)—provides the best predictions and how certain we can be about this. We then re-estimate each model using a one-third holdout sample, which we use to compare each model’s out-of-sample hit rate for churn, the Gini coefficient, and retention and the mean days’ error in terms of churn (i.e., how many days we miscalculate the moment of churn). We do this to detect which data source has the highest value for predicting churn and whether combining data sources results in incremental value.

Next, to uncover which (combination of) data source(s) best predicts churn, we also evaluate the (incremental) monetary value of each data source, to verify whether it is actually worthwhile for firms to collect and analyze all these data. To do so, we calculate the costs of misclassification per model per 1000 customers, using each model’s hit rates for churn and retention. From our data provider, we received the following information regarding the costs of misclassifying retainers and churners (based on 2015 data):

- The average cost for falsely classifying someone as a retainer is $170. This is based on the average likelihood of saving a correctly classified churner and the potential savings in CLV.
- The average cost for falsely classifying a customer as a churner is $40. This cost is the offered discount for customers who are predicted to churn to prevent them from churning.
• The industry average churn rate of 11%. We use this rate, which is similar to the annual churn rate at our data provider, to calculate the total misclassification costs.

**RESULTS**

Table 6 provides an overview of the parameter estimates of the eight estimated models, plus their fit criteria. If we analyze the three models using the individual data sources separately (i.e., Models 2–4 in Table 6), we find that each performs better than the baseline model on all fit criteria. The independent value of the variables from the customer database is $10,105 per 1000 customers, the NPS has a value of $8,180 per 1000 customers, and the textual data has a value of $4,759. Among the three data sources, the data from the customer database performs best on all fit criteria.

Combining two data sources, as in Models 5–7 in Table 6, improves the models, compared with the respective models with only one of the data sources used, according to most fit criteria. The best model, based on all fit criteria and the improvement in reducing the costs of misclassification, is the model using all three data sources. Given the Akaike weights, we can be certain (i.e., close to 100%, as indicated by the Akaike weight of 1.000) that the full model (i.e., using all information) provides the best predictions.

When calculating the incremental monetary value of each individual data source, we observe that for the model that does not take the textual data into account (i.e., Model 5), the total cost of misclassification is $28,915. This reduces to $27,656 when we also include the new variables from the textual data (Model 8), meaning that these data have an incremental monetary value of ($28,915 – $27,656) $1,259. Similarly, the data from the customer database have an incremental monetary value of ($34,894 – $27,656) $7,238, and the NPS has an incremental monetary value of ($30,921 – $27,656) $3,265.
Firms can use this information to investigate whether collecting, storing, and analyzing such data is worthwhile (e.g., a positive return on investment). In our specific case, an incremental monetary value of $1,259 per 1000 customers of the textual data means that our data provider, in the one country from which the data set comes and for the average number of customers who responded to the questionnaire, would attain an annual savings rate of approximately $315,000. Thus, with these data, firms can decide whether the potential savings are outweighed by the additional costs of handling these data. (When data are already collected and the process is automated, the incremental costs can be lower.)

When comparing the fit criteria with previous studies (e.g., De Haan, Verhoef, and Wiesel 2015), we find that the hit rate for the best-performing model of 65.1% is somewhat better than our churn hit rate of 62.9%, but substantially lower than the retention hit rate of 82.8%. Thus, the additional data do seem valuable for predicting churn and retention. Similarly, the Gini coefficient of our best model of .358 is substantially better compared to the average Gini coefficient .269 found by Neslin et al. (2006) and close to the best performing model found in their paper, which has a Gini coefficient of .41. We can thus conclude that the models which we have estimated are in general well able to predict churn.

When examining the individual parameters of the full model (Model 8), we show that the customers who have been with the firm for a longer time before calling the customer service center are less likely to churn. Surprisingly, waiting longer in the queue is negatively related to the likelihood to churn. An explanation might be related to the time of day; in addition, the negativity of waiting longer is already captured in the NPS and the polarity score. A longer conversation time is also related to a lower likelihood to churn, possibly because a longer conversation time can indicate that the service employee spent more time helping the customer or that the specific problem took more time to be resolved. The NPS is negatively related to churn, which we expected and is in line with previous studies (e.g., De
Haan, Verhoef, and Wiesel 2015). Four out of the six topics are all significantly related to churn; that is, they differ in the likelihood to churn compared with the baseline (i.e., no identified topic). The polarity (i.e., the sentiment of the text) is negatively and significantly related to churn; this means that, when we control for all other factors, the more positive the sentiment of the written feedback, the less likely the customer will churn after having contacted the customer service center.

As Table 6 shows, some of the parameters are time dependent (i.e., they become more or less strongly related to the probability to churn the more time has passed since the service call). Figure 2 shows two examples of the change of the parameters over time—for ln(Years with the firm) and NPS. As we show, ln(Years with the firm) right after the call is less negatively related to churn, while the more time has passed since the service call, the stronger is the relationship between ln(Years with the firm) and churn. A possible explanation is that a service call is a critical moment in the customer experience for both long-term and recent customers; thus, the length of the relationship matters less around these critical moments than in periods when there are no such critical moments.

For the NPS, we show a different pattern. Just after the service call, the NPS is strongly related to churn, while it becomes less important and comes close to being nonsignificant after about half a year. A possible explanation is that the NPS indicates something about the customer’s satisfaction with the service received from the customer service operator, which can have an impact on his or her decision to stay or leave directly after the experience, but it does not influence the behavior much more when time has passed. This pattern can also indicate something about the usability of such metrics: if a firm asked the customer a longer time ago, the NPS and generally CFMs should no longer be used. Thus, these metrics should be measured at least once every few months to make meaningful predictions.
CUSTOMER DASHBOARD

In addition to using polarity scores to predict the likelihood of churn, we can use the information extracted from unstructured textual data within a firm’s customer dashboard. Customer dashboards, especially those including mindset metrics, are useful to predict various financial and market outcomes, not only for customers providing these metrics, but across the whole customer base (e.g., Srinivasan and Hanssens 2009; Srinivasan, Vanhuele, and Pauwels 2010). These customer dashboards typically track information collected through surveys and are monitored over time using periodic (e.g., daily, weekly, monthly) information. Combined with information collected through surveys, the information extracted from unstructured data can also be incorporated into a customer dashboard.

For this purpose, firms can investigate how the sentiment and other characteristics of a text develop over time; how the characteristics of the texts differ between segments of customers, between regions, and so on; and the effect of all this information on (future) customer behavior and firm performance, this also in order to extrapolate the information from the (relatively small) group of customers providing textual feedback to the entire customer base. Figure 3 depicts a simple example of such a dashboard. Here, we compare the percentage of customers churning within 28 days after contacting the customer service center with the net polarity (i.e., the percentage of feedback with a positive sentiment less the percentage of feedback with a negative sentiment), and three of the emotions from the written feedback (derived from the textual feedback in the R-package “syuzhet”; for more details, see Jockers 2013).

We show that all the information extracted from textual data is tied to the churn rate within the next 28 days (this could be an early indicator or a warning sign for managers that action is required). For example, if a manager observes a drop in the polarity score (e.g., an
increase in the word “anger” or a decrease in the words “joy” or “trust” in the written feedback), he or she should take immediate action; otherwise, churn will likely increase.

Firms could thus create a dashboard based on unstructured data and connect this with (time-related) important outcome variables and other key performance indicators. In addition to customer-level outcomes (e.g., retention), dashboards can be used for firm-level (financial) outcome variables and can be tied to other (e.g., marketing-related) input variables (see Srinivasan and Hanssens 2009; Srinivasan, Vanhuele, and Pauwels 2010). In this sense, unstructured data can be an addition or even an alternative to (expensive) survey-based metrics.

Alternatively it can also be investigated what is driving certain written comments, and who are responsible for what kind of feedback. Since not all customers contact the service center and not everyone who contacts the service center provides a (written) feedback, dashboards enable to find out what opinions of other (silent) customers might be. For example, Table 7 shows how the CRM-type variables relate to the topic and polarity of the written feedback. For this we have used both a multinomial logistic regression model (model 1 in Table 7) to explain topic membership and an OLS regression model (model 2 in Table 7) to explain the variance in the polarity score.

What we can for instance observe from the results in Table 7 is that older customers and males are less likely to mention a topic, i.e. all parameters for these groups in model 1 in table 7 are significant and negative, meaning these customers are more likely to have written a comment related to the reference category (‘no topic’). Customer who had a longer relationship to the firm had a higher likelihood of writing a topic. Understanding who is more likely to write a topic can help understand which customers are more/less likely to write comments, which can in turn help to understand how generalizable the information is for the whole customer base and to what extent it can(not) be extrapolated to other customers.
Customers who were longer in the waiting queue write more comments about either the call or the slow service, in line with what we might suspect. The same topics are brought up by customers who have been called back. All factors in the model also significantly relate to the polarity score; older females, who have been longer with the firm, were less long in the waiting queue, had a longer conversation, and had called before write a more positive comment. Customers who are called back are overall less positive; this might actually be the reason why they are called back, i.e. the firm trying to improve or restore the relationship with the customer.

All of this can be used to better understand individual customers, both those who have provided written feedback, as well as those who haven’t done so. The regression outcomes from Table 7 can furthermore be used to correct for non-representativeness of customers who do contact the customers service center and who do after this provide written feedback, e.g. to provide more representative information for the customer dashboard and to help improve overall customer service.

**CONCLUSION AND MANAGERIAL IMPLICATIONS**

Many firms now have access to a large amount of unstructured data. Although these rich data can provide better and deeper insights into the customer base, they are usually not considered in CRM, mainly because of their nature and the difficulty of directly combining them with structured data from the database. In our study, we aimed to overcome this barrier by showing how to extract meaningful insights from unstructured data and how to include these in econometric models. We empirically showed that incorporating such data in the model significantly improves the estimation of churn and retention, compared with models that use only traditional variables. In our study case, given the better prediction of churn and retention, the data provider can better target potential churners, which according to its figures can lead to an annual savings of $315,000. These potential savings though will also depend on
the exact data sources and variables available, the type and volume of unstructured data, the
element to be predicted, the number of customers, the type of industry, and so on. Given that
firms have a great deal of unstructured data available, the methods and simulation used in our
study can easily be applied to other settings, such that for each situation, firms can calculate
whether collecting and analyzing these data is worthwhile (i.e., if the gains outweigh the
costs).

Furthermore, with text mining, written customer feedback can provide extensive
insights that firms can also incorporate into their customer dashboards, to gain a better
understanding of the development of customer attitudes over time. This gain is not accessible
through simply the raw written feedback, because each written feedback is unique and its
hand-operated examination can be laborious and too expensive.

Firms can alternatively investigate some of the assimilated elements, such as
emotions, using surveys (e.g., Ou and Verhoef 2017), though structured surveys have a
disadvantage in that respondents can only give feedback on the elements in the questionnaire.
Proposing open-ended questions and analyzing these with text-mining methods, as we did in
our study, can uncover other potential issues not included in structured questionnaires.

This research has several limitations that might lead to further research. First, the type
of unstructured data (written feedback) available for our case study is relatively scant,
provided by customers after they contacted the customer service center. Many other data
sources in practice might be available, including firm–customer interactions (e.g., through
e-mail and chats), written comments on social media and review sites, and other forms of
multimedia such as voice (e.g., spoken conversations) and video (e.g., video reviews on
YouTube). Thus, it would be worthwhile to consider various data sources and forms of
unstructured data, which might lead to recommendations for the use of different analytical
tools and thus provide novel insights for firms.
Second, we made use of non-English written feedback. A downside of this is that many sophisticated tools such as sentiment analysis are not always available or as good for other languages. We overcame this limitation by translating our text into English. In cases in which the original text is in English or the tools are more advanced or specifically developed for a particular language, the results could potentially improve. Consequently, our results may be on the conservative side in terms of predicting the value of these data.

Third, we examined only customer retention. We selected this outcome variable because it is the most important component of CLV, which in turn is strongly related to firm value (Gupta, Lehmann, and Stuart 2004). However, research could use other relevant (customer- and firm-level) outcome variables, such as profitability, cross-buying, upselling, and word of mouth, as objects of prediction. If customers’ written feedback can also help explain and predict such variables, the value of these data would be even higher (i.e., assessing only retention provides only a conservative estimate of the value of these data).

Finally, our study is rather descriptive in nature. We observe written feedbacks at one period in time and use the insights from these data to make (in- and out-of-sample) predictions about future customer behavior. However, this does not necessarily entail causal relations; that is, we cannot conclude that improving the sentiment of the written feedback (by leveraging, for example, the service level of the call center) will lead to a subsequent improvement in retention. Thus, further research can focus on this.
REFERENCES


Hu, Minqing and Bing Liu (2004), “Mining and summarizing customer reviews,”
Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and
Data Mining, August, 168-177.

Indicators of Financial Performance? An Analysis of Customer Satisfaction,” Journal
of Accounting Research, 36 (3), 1-35.

Jiang, Zhenchao, Lishuang Li, and Degen Huang (2016), “An Unsupervised Graph Based
Continuous Word Representation Method for Biomedical Text Mining,” IEEE/ACM
Transactions on Computational Biology and Bioinformatics, 13 (4), 634-42.

project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

Keiningham, Timothy L., Bruce Cooil, Lerzan Aksoy, Tor W. Andreassen, and Jay Weiner
Predicting Customer Retention, Recommendation, and Share-of-Wallet,” Managing
Service Quality, 17 (4), 361-84.

Keiningham, Timothy L., Bruce Cooil, Tor Wallin Andreassen, and Lerzan Aksoy (2007), “A
Longitudinal Examination of Net Promoter and Firm Revenue Growth,” Journal of

NY.

Kumar, V. and Werner Reinartz (2016), “Creating Enduring Customer Value,” Journal of
Marketing, 80 (6), 36-68.

Lemmens, Aurélie, and Christophe Croux (2006), “Bagging and Boosting Classification Trees
to Predict Churn,” Journal of Marketing Research, 43 (2), 276-86.


<table>
<thead>
<tr>
<th></th>
<th>Customer Database Data</th>
<th>Survey Data</th>
<th>Unstructured Data</th>
<th>Time Dependency</th>
<th>Method</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lemmens and Croux</strong></td>
<td>Psychographics,</td>
<td>N.A.</td>
<td>N.A.</td>
<td>No</td>
<td>Classification trees</td>
<td>Churn</td>
</tr>
<tr>
<td><strong>(2006)</strong></td>
<td>demographics, RFM,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>product specifics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risselada, Verhoef</strong></td>
<td>Demographics, RFM,</td>
<td>N.A.</td>
<td>N.A.</td>
<td>No</td>
<td>Logistic regressions,</td>
<td>Churn</td>
</tr>
<tr>
<td><strong>and Bijmolt (2010)</strong></td>
<td>product specifics,</td>
<td></td>
<td></td>
<td></td>
<td>classif, trees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relationship length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mittal and Kamakura</strong></td>
<td>Demographics</td>
<td>CS</td>
<td>N.A.</td>
<td>No</td>
<td>Regression models</td>
<td>Repurchase</td>
</tr>
<tr>
<td><strong>(2001)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>De Haan, Verhoef,</strong></td>
<td>Demographics,</td>
<td>CS, NPS,</td>
<td>N.A.</td>
<td>No</td>
<td>Probit models</td>
<td>Churn</td>
</tr>
<tr>
<td><strong>and Wiesel (2015)</strong></td>
<td>relationship length</td>
<td>CES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tirunillai and Tellis</strong></td>
<td>N.A.</td>
<td>N.A.</td>
<td>Positive vs.</td>
<td>No</td>
<td>VAR</td>
<td>Abnormal stock</td>
</tr>
<tr>
<td><strong>(2012)</strong></td>
<td></td>
<td></td>
<td>negative online</td>
<td></td>
<td></td>
<td>return and trade</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chatter</td>
<td></td>
<td></td>
<td>volume</td>
</tr>
<tr>
<td><strong>Salehan and Kim</strong></td>
<td>N.A.</td>
<td>N.A.</td>
<td>Sentiment of</td>
<td>No</td>
<td>Regression models</td>
<td>Readership +</td>
</tr>
<tr>
<td><strong>(2016)</strong></td>
<td></td>
<td></td>
<td>online review</td>
<td></td>
<td></td>
<td>helpfulness</td>
</tr>
<tr>
<td><strong>Zhang, Moe,</strong></td>
<td>N.A.</td>
<td>N.A.</td>
<td>Message (tweet)</td>
<td>No</td>
<td>Split hazard process</td>
<td>Retweeting</td>
</tr>
<tr>
<td><strong>and Schweidel (2017)</strong></td>
<td></td>
<td></td>
<td>content</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Netzer, Lemaire and</strong></td>
<td>Demographics,</td>
<td>N.A.</td>
<td>Content in loan</td>
<td>No</td>
<td>Logistic regressions,</td>
<td>Loan default</td>
</tr>
<tr>
<td><strong>Herzenstein (2017)</strong></td>
<td>financial information</td>
<td></td>
<td>applications</td>
<td></td>
<td>trees</td>
<td></td>
</tr>
<tr>
<td><strong>Coussement and Van</strong></td>
<td>Demographics,</td>
<td>N.A.</td>
<td>Positive vs.</td>
<td>No</td>
<td>Split hazard process</td>
<td></td>
</tr>
<tr>
<td><strong>den Poel (2009)</strong></td>
<td>transactional</td>
<td></td>
<td>negative emotions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>information, RFM</td>
<td></td>
<td>in compliant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>emails and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>requests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This study</td>
<td>Demographics, relationship length, service interaction contact details</td>
<td>NPS</td>
<td>Sentiment and content of service call feedback</td>
<td>Yes</td>
<td>Cox prop. hazard model with time-dep. covariates</td>
<td>Churn</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>-----</td>
<td>-----------------------------------------------</td>
<td>-----</td>
<td>-------------------------------------------------</td>
<td>-------</td>
</tr>
</tbody>
</table>

Notes: CS = customer satisfaction, VAR = vector autoregression model, SVM = support vector machines.
Table 2
Descriptive Statistics (n = 44,192).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Churn</td>
<td>43.6% of customers churned before February 16, 2015, 56.4% retained</td>
</tr>
<tr>
<td>Gender</td>
<td>Male (53.7%), female (46.3%)</td>
</tr>
<tr>
<td>Product category</td>
<td>Broadband fiber (.38%), Broadband xDSL (4.8%), ‘youth’ (18.4%), mobile (62.4%),</td>
</tr>
<tr>
<td></td>
<td>mobile broadband (1.7%), other (.3%), unknown (12.0%)</td>
</tr>
<tr>
<td>Time of day</td>
<td>Early morning (27.2%), later morning (20.4%), early afternoon (29.3%),</td>
</tr>
<tr>
<td></td>
<td>later afternoon/evening (23.1%)</td>
</tr>
<tr>
<td>Day of the week</td>
<td>Monday (18.8%), Tuesday (18.1%), Wednesday (17.3%), Thursday (16.9%), Friday</td>
</tr>
<tr>
<td></td>
<td>(17.4%), Saturday (7.2%), Sunday (4.3%)</td>
</tr>
<tr>
<td>Month of year</td>
<td>Jan. (8.4%), Feb. (5.2%), Mar. (8.2%) Apr. (7.4%), May (8.3%), June (10.5%),</td>
</tr>
<tr>
<td></td>
<td>July (9.4%), Aug. (6.4%), Sep. (6.3%), Oct. (12.7%), Nov. (9.1%), Dec. (8.1%)</td>
</tr>
<tr>
<td>Year</td>
<td>2013 (37.0%), 2014 (63.0%)</td>
</tr>
<tr>
<td>Repeated_call</td>
<td>Yes (95.0%), No (5.0%)</td>
</tr>
<tr>
<td>Called_back</td>
<td>Yes (16.3%), No (83.7%)</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td><strong>SD</strong></td>
</tr>
<tr>
<td>Years_at_firm</td>
<td>5.50</td>
</tr>
<tr>
<td>Age</td>
<td>51.09</td>
</tr>
<tr>
<td>Minutes_in_queue</td>
<td>.99</td>
</tr>
<tr>
<td>Minutes_conversation</td>
<td>5.44</td>
</tr>
<tr>
<td>NPS</td>
<td>7.27</td>
</tr>
</tbody>
</table>
Table 3  
Correlation Matrix (n = 44,192).

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Churn</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Gender (female)</td>
<td>-.027**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Repeated_call</td>
<td>.017**</td>
<td>.017**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) Called_back</td>
<td>-.052**</td>
<td>.004ns</td>
<td>.001ns</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) Years_at_firm</td>
<td>-.362**</td>
<td>-.017**</td>
<td>-.007ns</td>
<td>.008ns</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6) Age</td>
<td>-.161**</td>
<td>-.033**</td>
<td>-.032**</td>
<td>.046**</td>
<td>.305**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7) Minutes_in_queue</td>
<td>-.016**</td>
<td>-.014**</td>
<td>.063**</td>
<td>-.070**</td>
<td>-.004ns</td>
<td>-.020**</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8) Minutes_conversation</td>
<td>-.056**</td>
<td>-.000ns</td>
<td>.034**</td>
<td>.031**</td>
<td>.053**</td>
<td>.072**</td>
<td>.044**</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(9) NPS</td>
<td>-.171**</td>
<td>.047**</td>
<td>.017**</td>
<td>-.033**</td>
<td>.092**</td>
<td>.103**</td>
<td>-.057**</td>
<td>.036**</td>
<td>1</td>
</tr>
</tbody>
</table>

** p < .01, * p < .05, ns p > .05.
Table 4
Example Polarity Score by qdap.

<table>
<thead>
<tr>
<th>Description</th>
<th>Polarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I am happy with the provided service.</td>
<td>.378</td>
</tr>
<tr>
<td>2. I am unhappy with the provided service.</td>
<td>-.378</td>
</tr>
<tr>
<td>3. I am not happy with the service.</td>
<td>-.378</td>
</tr>
<tr>
<td>4. I am barely happy with the service.</td>
<td>.076</td>
</tr>
<tr>
<td>5. I am very happy with the service.</td>
<td>.680</td>
</tr>
<tr>
<td>6. I am very unhappy with the service.</td>
<td>-.680</td>
</tr>
<tr>
<td>7. I am not happy with the provided service.</td>
<td>-.354</td>
</tr>
<tr>
<td>8. I am barely happy with the provided service.</td>
<td>.071</td>
</tr>
<tr>
<td>9. I am very happy with the provided service.</td>
<td>.636</td>
</tr>
<tr>
<td>10. I am very unhappy with the provided service.</td>
<td>-.636</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Size %</td>
<td>16.6%</td>
</tr>
<tr>
<td>#1 word</td>
<td>Ring</td>
</tr>
<tr>
<td>#2 word</td>
<td>Time</td>
</tr>
<tr>
<td>#3 word</td>
<td>Calling</td>
</tr>
<tr>
<td>#4 word</td>
<td>Telephone</td>
</tr>
<tr>
<td>#5 word</td>
<td>Day</td>
</tr>
<tr>
<td>Pos. polarity</td>
<td>32.0%</td>
</tr>
<tr>
<td>Neg. polarity</td>
<td>38.4%</td>
</tr>
<tr>
<td>No polarity</td>
<td>29.6%</td>
</tr>
<tr>
<td>Churn</td>
<td>49.1%</td>
</tr>
</tbody>
</table>
Table 6
Model Comparison (n = 44,192).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.001 n.s.</td>
<td></td>
<td></td>
<td>.001 n.s.</td>
<td>-.001 n.s.</td>
<td>.001 *</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.058 **</td>
<td></td>
<td></td>
<td>.038 **</td>
<td>.048 **</td>
<td>.037 **</td>
<td></td>
</tr>
<tr>
<td>Ln(Years with the firm)</td>
<td>-1.336 **</td>
<td></td>
<td></td>
<td>-1.314 **</td>
<td>-1.326 **</td>
<td>-1.316 **</td>
<td></td>
</tr>
<tr>
<td>Ln(Time in queue)</td>
<td>-.019 n.s.</td>
<td></td>
<td></td>
<td>-.036 **</td>
<td>-.036 *</td>
<td>-.035 *</td>
<td></td>
</tr>
<tr>
<td>Ln(Conversation minutes)</td>
<td>-.055 **</td>
<td></td>
<td></td>
<td>-.048 **</td>
<td>-.051 ***</td>
<td>-.048 ***</td>
<td></td>
</tr>
<tr>
<td>NPS</td>
<td>-.085 ***</td>
<td>-0.065 **</td>
<td></td>
<td>-.0812 ***</td>
<td>-.062 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Call(ing) related</td>
<td></td>
<td></td>
<td></td>
<td>1.223 **</td>
<td>.398 n.s.</td>
<td>-1.154 **</td>
<td>-1.404 **</td>
</tr>
<tr>
<td>2. Customer service</td>
<td></td>
<td></td>
<td></td>
<td>-1.692 **</td>
<td>-1.323 **</td>
<td>-.233 n.s.</td>
<td>-.207 n.s.</td>
</tr>
<tr>
<td>3. Speed/coverage</td>
<td></td>
<td></td>
<td></td>
<td>-1.317 ***</td>
<td>-1.402 ***</td>
<td>.447 n.s.</td>
<td>-.031 n.s.</td>
</tr>
<tr>
<td>4. Quick help</td>
<td></td>
<td></td>
<td></td>
<td>-1.871 ***</td>
<td>-.971 **</td>
<td>-1.568 **</td>
<td>-.750 **</td>
</tr>
<tr>
<td>5. Subscription related</td>
<td></td>
<td></td>
<td></td>
<td>3.306 ***</td>
<td>3.831 ***</td>
<td>2.621 ***</td>
<td>3.264 ***</td>
</tr>
<tr>
<td>6. Slow service</td>
<td></td>
<td></td>
<td></td>
<td>-.087 n.s.</td>
<td>-.046 n.s.</td>
<td>-.238 n.s.</td>
<td>-0.696 *</td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
<td></td>
<td>-.203 **</td>
<td>-0.163 ***</td>
<td>-.038 ***</td>
<td>-0.036 ***</td>
</tr>
<tr>
<td>Control for product type</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Control for time of day</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Control for day of the week</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Control for month</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Control for year</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>392367</td>
<td>379369</td>
<td>390646</td>
<td>391529</td>
<td>378362</td>
<td>378830</td>
<td>390295</td>
</tr>
<tr>
<td>BIC</td>
<td>392375</td>
<td>379637</td>
<td>390654</td>
<td>391584</td>
<td>378638</td>
<td>379152</td>
<td>390398</td>
</tr>
<tr>
<td>Akaike weights</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Hit-rate churn</td>
<td>40.2%</td>
<td>60.1%</td>
<td>47.3%</td>
<td>45.2%</td>
<td>61.8%</td>
<td>60.9%</td>
<td>49.7%</td>
</tr>
<tr>
<td>Hit-rate retention</td>
<td>73.3%</td>
<td>79.4%</td>
<td>78.5%</td>
<td>76.3%</td>
<td>82.0%</td>
<td>80.7%</td>
<td>78.4%</td>
</tr>
<tr>
<td>Gini coefficient</td>
<td>.135</td>
<td>.313</td>
<td>.179</td>
<td>.169</td>
<td>.323</td>
<td>.321</td>
<td>.165</td>
</tr>
<tr>
<td>Mean days off</td>
<td>275.6</td>
<td>176.3</td>
<td>196.7</td>
<td>194.9</td>
<td>174.6</td>
<td>175.7</td>
<td>194.0</td>
</tr>
<tr>
<td></td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Cost of churn misc.¹</td>
<td>2,631</td>
<td>1,756</td>
<td>2,319</td>
<td>2,411</td>
<td>1,681</td>
<td>1,720</td>
<td>2,213</td>
</tr>
<tr>
<td>Cost of retention misc.</td>
<td>40,397</td>
<td>31,168</td>
<td>32,530</td>
<td>35,858</td>
<td>27,234</td>
<td>29,201</td>
<td>32,681</td>
</tr>
<tr>
<td>Total costs of misc.</td>
<td>43,028</td>
<td>32,923</td>
<td>34,848</td>
<td>38,269</td>
<td>28,915</td>
<td>30,921</td>
<td>34,894</td>
</tr>
<tr>
<td>Improvement from base</td>
<td>$</td>
<td>10,105</td>
<td>8,180</td>
<td>4,759</td>
<td>14,114</td>
<td>12,107</td>
<td>8,134</td>
</tr>
</tbody>
</table>

*** p < .01, ** p < .05, * p < .10, n.s. p > .10. Parameters are in italics; bold and underlined are time dependent.

¹ Per 1000 customers. For the full model, 37.1% of churners are misclassified (62.9% are correctly classified), with a mean churn rate of 11% and $40 costs of misclassifying churners. This results in a cost of .371 × $40 × .11 × 1000 customers = $1,632. The cost of retention misclassification is calculated in the same way.
Table 7
Prediction of textual topic (model 1) and polarity (model 2) (n = 44,192).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.743***</td>
<td>.879***</td>
<td>1.569***</td>
<td>.892***</td>
<td>.441ns.</td>
<td>1.476***</td>
<td>.164***</td>
</tr>
<tr>
<td>Age</td>
<td>-.011***</td>
<td>-.008***</td>
<td>-.013***</td>
<td>-.005***</td>
<td>-.011***</td>
<td>-.010***</td>
<td>.001***</td>
</tr>
<tr>
<td>Male</td>
<td>-.343***</td>
<td>-.587***</td>
<td>-.234***</td>
<td>-.362***</td>
<td>-.293***</td>
<td>-.237***</td>
<td>-.040***</td>
</tr>
<tr>
<td>Ln(Years with the firm)</td>
<td>.071**</td>
<td>.239***</td>
<td>.242***</td>
<td>.232***</td>
<td>.221***</td>
<td>.104***</td>
<td>.042***</td>
</tr>
<tr>
<td>Ln(Time in queue)</td>
<td>.087**</td>
<td>.003ns.</td>
<td>-.035ns.</td>
<td>.016ns.</td>
<td>-.031ns.</td>
<td>.359***</td>
<td>-.074***</td>
</tr>
<tr>
<td>Ln(Conversation minutes)</td>
<td>.181***</td>
<td>.368***</td>
<td>.186***</td>
<td>-.009ns.</td>
<td>.122***</td>
<td>.123***</td>
<td>.024***</td>
</tr>
<tr>
<td>Repeated Call</td>
<td>-.081ns.</td>
<td>.187*</td>
<td>.192*</td>
<td>.250**</td>
<td>.070ns.</td>
<td>-.285ns.</td>
<td>.072***</td>
</tr>
<tr>
<td>Has been called back</td>
<td>.154**</td>
<td>-.041ns.</td>
<td>.005ns.</td>
<td>-.055ns.</td>
<td>.096ns.</td>
<td>.193**</td>
<td>-.057***</td>
</tr>
</tbody>
</table>

Control for product type: √
Control for time of day: √
Control for day of the week: √
Control for month: √
Control for year: √

*** p < .01, ** p < .05, * p < .10, ns p > .10.
Figure 1
Histogram of Polarity Score (n = 44,192).
Figure 2
Varying Parameter Estimates Over Time

Notes: The dotted line indicates the mean parameter estimate, while the curved line shows that this parameter is dependent on the time since the service call (including a 95% confidence interval).
Figure 3
Example of an Unstructured Data-Based Dashboard