An Analysis of Firm Related Communications using a Marketing Lens

Dissertation

Submitted as part of the requirements under the

Fellow Program in Management, Indian School of Business

Submitted by

Ashish S. Galande (51250002)

FPM Candidate in Marketing

Chair
Professor Sudhir Voleti

Committee Members
Professor Siddharth S. Singh
Professor Sundar G. Bhardwaj

July 2018

Abstract

The determinants of firm performance have elicited a lot of research attention, particularly in the economics and management literatures. A generally acknowledged perspective is that firm outcomes are a certain function of (i) internal factors like firm demographics, organizational structure, capital structure, (ii) of external factors like industry structure and characteristics, regulatory environment, macroeconomic conditions, and (iii) of firm strategy, which connects these internal and external factors. However, firm strategy is an intangible quantity whose measurement, modeling and analysis have proven to be a challenge. Managers, when asked to describe firm strategy, do so in words. Most external communication by firms, used for informing various stakeholders the firms' strategy, is textual in form. Firms' compliance filings, made at the behest of the government and regulators, often contain strategic content of interest and are again primarily textual in form. At the same time, most individuals and groups communicate by writing (which is textual in nature) or speech (transcribe-able to text). Consequently, the analysis of text data has assumed significance in various fields, especially in the social sciences. Text data (e.g. speech or interview transcripts, complaint emails, offer documents and prospectuses, product reviews, etc.) often capture qualitative aspects of information on a subject that may otherwise be unavailable. My research questions are motivated by these communications originating from or are about a firm. I infer the latent information, which has been studied in the past as "voice" or "mindset", using text analytic methods and study its impact on firm performance.

I classify all firm related communications into two categories – one that originates from the firm and the other that originates from any of its stakeholders (e.g. consumers, analysts, expert reviewers). The other dimension to these communications is the context in which they exist, i.e. is it a statutory or solicited form of communication (e.g. SEC filings, solicited open-ended responses

from consumers) or is it voluntary (e.g. press release, letter to the shareholders, expert reviews, blogs). This classification allows us to *a priori* classify the communications into groups based on their origin and their nature. Having classified the forms of communications into these four quadrants, I then draw upon extant text-analytic methods to mine and extract qualitative themes or *topics* of interest underlying a mass of raw text responses, and develop a scalable approach to test these subjective themes for explanatory and predictive validity of firm performance.

In this dissertation I use a marketing lens to measure the firm related communications from unstructured data to better understand its implications for consumers, firms and regulators. I define firm related communications broadly to include all types of discussion captured in unstructured format- consumer opinions as captured in consumer reviews and blogs, intra firm discussions and, public communications by firms and regulatory institutions. I combine econometric methods and natural language technologies to examine the influences on and of firm related communications. The first of the three essays proposes and validates an extension to a popular latent topic modeling algorithm which enables better interpretation of consumer insight from these communications. The second essay uses this proposed extension derive a measure of marketing effectiveness from communications and uses it to explain brand performance. The third essay uses this proposed extension to classify firms as following a niche or popular product-market choices based on strategic information in their statutory filings. I find a statistically significant performance differential attributable to product-market choices.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am truly fortunate to have met my doctoral advisor, Professor Sudhir Voleti and the committee members Professor Siddharth S. Singh and Professor Sundar G. Bharadwaj. I am indebted to them, who have spent countless hours guiding the work presented here. It was a great honor for me to have the opportunity of working under them.

Professor Voleti was a figure of the true scholar that I had always imagined during my entire academic path before I met him. I learned how difficult and how important doing research in the right way is. His unlimited passion and curiosity for academic research pushed me to want myself to be a better researcher. The lessons from our meetings will be an unextinguishable lighthouse in my future academic voyage after ISB. His patience and generosity made this dissertation, which seemed impossible at many times, possible.

Professor Singh formed me into a better balanced researcher. His approach to research encouraged me to clearly know the strengths and weaknesses in my ability as a researcher. He was instrumental in encouraging me to consider alternative means of collecting data when negotiations with one of our data providers had not succeeded as planned.

Professor Bharadwaj offered not only selfless guidance, but also the encouragement that helped me stay motivated during my doctoral studies. His support and guidance were always unconditional. His vast knowledge and experience over research and practice enhanced my strengths to a higher level and my weaknesses supplemented. I hope that we can work on more exciting research projects together.

I also want to acknowledge the guidance provided by the resident and the visiting faculty at ISB. They were always open to answer random questions and provided insightful comments on this dissertation as well as my early pieces of research. I truly appreciate the thoughtful insights they have provided me.

Next, I would like to thank the program office who helped me navigate my way through the program until the end and for providing unconditional emotional support. I am very grateful to all the fellow students and research associates who have doubled up as colleagues and critics.

This dissertation would not have been possible without the unconditional support of my family and Prakash Satyavageeswaran, all of whom have supported me in every possible way.

Thank you all very much.

Chapter 1 - Introduction

The determinants of firm performance have elicited a lot of research attention, particularly in the economics and management literatures. A widely accepted perspective is that firm performance is some function of factors internal to the firm (e.g., firm demographics, organizational structure, capital structure etc.), of factors external to the firm such as environmental variables (e.g., industry structure and characteristics, regulatory environment, macroeconomic conditions etc.), and of firm strategy, which connects and reconciles these internal and external factors. However, firm strategy is an intangible quantity whose measurement, modeling and analysis have proven to be a challenge. Managers, when asked to describe firm strategy, do so in words. Most external communication by firms, used for informing various stakeholders the firms' strategy, is textual in form. Firms' compliance filings, made at the behest of the government and regulators, often contain strategic content of interest and are again primarily textual in form. At the same time, most individuals and groups communicate by writing (which is textual in nature) or speech (transcribe-able to text). Consequently, the analysis of text data has assumed significance in various fields, especially in the social sciences. Text data (e.g. speech or interview transcripts, complaint emails, offer documents and prospectuses, product reviews, etc.) often capture qualitative aspects of information on a subject that may otherwise be unavailable. My research questions are motivated by these communications originating from or are about a firm. I infer the latent information, which has been studied in the past as "voice" or "mindset", using text analytic methods and study its impact on firm performance.

I classify all firm related communications into two categories – one that originates from the firm and the other that originates from any of its stakeholders (e.g. consumers, analysts, expert

reviewers). The other dimension to these communications is the context in which they exist, i.e. is it a statutory or solicited form of communication (e.g. SEC filings, solicited open-ended responses from consumers) or is it voluntary (e.g. press release, letter to the shareholders, expert reviews, blogs). This classification allows us to *a priori* classify the communications into groups based on their origin and their nature, as shown in Figure 1.1. Having classified the forms of communications into these four quadrants, I then draw upon extant text-analytic methods to mine and extract qualitative themes or *topics* of interest underlying a mass of raw text responses, and develop a scalable approach to test these subjective themes for explanatory and predictive validity of firm performance.

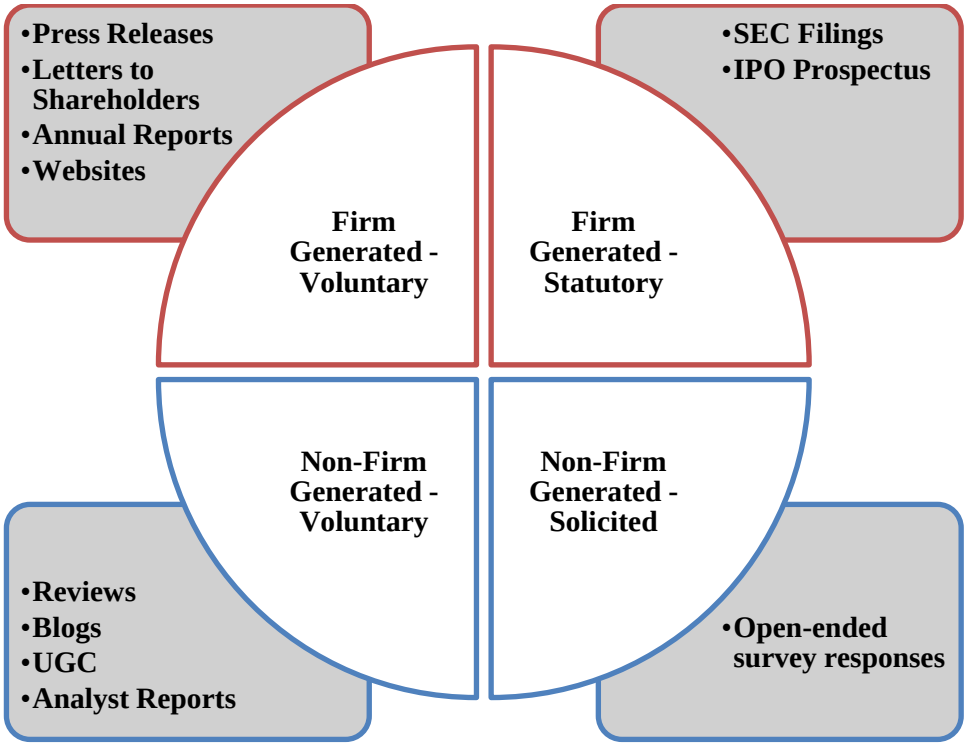


Figure 1.1 Firm related communications

In this dissertation I use a marketing lens to measure the firm related communications from unstructured data to better understand its implications for consumers, firms and regulators. I define

firm related communications broadly to include all types of discussion captured in unstructured format- consumer opinions as captured in consumer reviews and blogs, intra firm discussions and, public communications by firms and regulatory institutions. I combine econometric methods and natural language technologies to examine the influences on and of firm related communications. The first of the three essays proposes and validates an extension to a popular latent topic modeling algorithm which enables better interpretation of consumer insight from these communications. The second essay uses this proposed extension derive a measure of marketing effectiveness from communications and uses it to explain brand valuation. The third essay uses the extension to arrive at latent product-market choices made by firms using unstructured textual data in regulatory filings. It then uses classifies firms as making niche or popular product – market choices and estimates the causal effect of the firms switching strategies over those firms which have maintained the same strategy. The abstracts of the three essays are provided below.

In the first essay (*Extending the LDA Approach for Improved Qualitative Analyses on Text Data*), I consider that a researcher often faces situations wherein interpretation of the topic of discussion in text documents is required. In this context the use of latent topic models have gained popularity. I consider the most widely used latent topic model - the Latent Dirichlet Allocation or LDA model and explore ways to mitigate limitations in topic interpretation and out-of-sample prediction that were concomitant with the conventional LDA method. I propose an approach based on post-processing of LDA output - the "extended" LDA - that provides one of the methods to sidestep these drawbacks. I assess the extended LDA model's topic interpretation and topic recovery performance in two ways: a simulation experiment and human-machine concordance studies. Finally, I empirically demonstrate applications of this approach using Amazon Product reviews and using 10-K statements of software and services firms.

The second essay (*Brand Marketing Effectiveness, Brand Positioning Gaps and Brand Outcomes: An Empirical Investigation*) proposes a measure of marketing effectiveness which is derived from the brand's own marketing communication and the communication content generated by consumers, media and other experts. The place that a brand occupies in the minds of the consumer such that it helps distinguish between products and competitors is defined as *positioning*. In this vein, a brand manager intends to position a brand with a certain understanding of the target consumer's mindset. The target consumer may or may not perceive the brand in a manner assumed by the brand manager. Using the "extended LDA" method proposed in essay 1, a latent topic analysis of brand's communication reveals the perception of the brand salient in the managers' mind. A similar latent topic analysis of the communication generated by consumers, media and other experts reveals the perception of the brand salient to them. This paper uses the gap in two perceptions to arrive at a measure of marketing effectiveness. Using the top 500 brands from brandirectory.com, I mine the textual communications from brand's website, expert blogs, newspaper and magazine articles, and social media to infer the two sets of perceptions. As proposed, I find that that marketing effectiveness as a measure derived from the *positioning gap* leads to higher brand valuation. The study provides brand managers with a tool to not only better understand their consumers but also provides them with a much-needed method to analyze the effectiveness of their communications.

The third essay (*The Dynamics of Product-Market Choices in the Technology Sector*) posits that marketing strategy salient in managers' mind - such as management's perceptions, priorities, plans, goals, constraints, strategies - are captured to a substantial extent in a firm's strategic communications. These considerations influence managerial decision making, which in turn drives firm behavior and thereby, firm outcomes. I use these communications to empirically

investigate the extent to which a firm's product-market choices inform market outcomes beyond those explained by observed firm characteristics, behavior and standard control factors. I leverage text analytic methods to classify firms as making a popular product-market choice or a niche choice. Further, firms may choose to switch strategies based on their assessment of the market situation and hence will expect a differential in performance outcomes. I endeavor to measure, model and estimate the effect of such strategic changes in firm behavior. The data comes from publicly traded firms in the Technology sector in the US. The product-market choices are mined from firms' 10-K filings with the SEC. The firm performance data comes from Compustat. I find that firms moving from a popular to niche product-market choice have a significantly larger effect on their performance when compared to those firms which have moved from niche to popular product-market choice or those firms which haven't changed their choices.

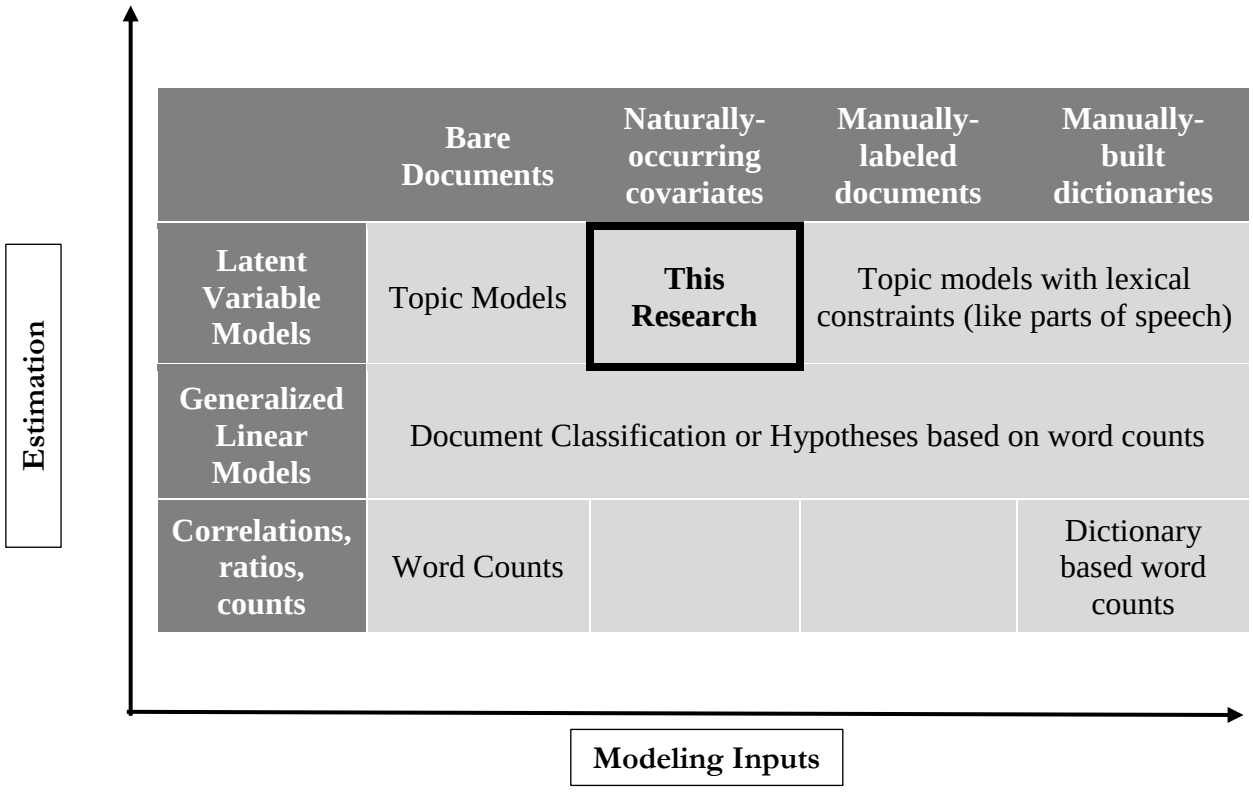
1.1. Positioning the current research

Techniques for automated content analysis of text are a combination of natural language processing, information retrieval, text mining, and machine learning. At the same time, the techniques employed range from simple analysis of comparative word frequencies to more complex hierarchical and mixture models. Figure 1.2 displays a sample of text analysis use in a variety of social science fields such as Political Science, Economics, Sociology, History and Management. Figure 1.3 attempts to position this study's contribution by locating it along two dimensions of interest - increasing estimation method complexity - from simple counts to generalized linear models (e.g., Weisberg 2005; Agresti & Kateri 2011; Gelman & Hill 2006; Elith et. al 2011 with logistic regressions) to latent variable methods (e.g., Blei et al. 2017; Mimno & McCallum 2012; Ramage et. al. 2011) - versus the increasing domain specificity of modeling

inputs (from raw text documents to manually labeled documents to custom-building semantic dictionaries). Each cell in the table represents the intersection of a certain type of estimation method and a certain type of data input. For instance, Pedregosa et al. (2011) use simple word frequency counts over year of book publication which would represent a naturally occurring covariate set. Other examples of model inputs include Manually-labeled documents such as annotated news articles describing rare events information (King and Lowe 2003 in the International relations literature) or semantic dictionaries, custom-built, reused or adapted from already-existing dictionaries (Freebase by Bollacker et.al. 2008 for names; LIWC by Tausczik & Pennebacker 2009 for affect).

- **Political Science:** Does social media reflect public political opinion, or forecast elections (O'Connor et al., 2010; Metaxas et al., 2011)? What determines international conflict and cooperation (Schrodt et al., 1994; King and Lowe, 2003; Shellman, 2008)?
- **Economics and Management**: How does sentiment in the media affect the stock market (Tetlock, 2007; Lavrenko et al., 2000)? Does sentiment in social media associate with stocks (Gilbert and Karahalios, 2010; Das and Chen, 2007; Bollen et al., 2011)? Do a company's SEC filings predict aspects of stock performance (Kogan et al., 2009; Loughran and McDonald, 2011)? What determines a customer's trust in an online merchant (Archak et al., 2011)?
- **Sociology of Science:** What are influential topics within a scientific community (Gerrish and Blei, 2010)?
- **Public Health:** How can search queries and social media help measure levels of the flu and better understand other public health issues (Ginsberg et al., 2009; Culotta, 2010; Paul and Dredze, 2011)?
- **History:** How did modern English legal institutions develop over the 17th to 20th centuries (Cohen et al., 2011)?

Figure 1.2 Some examples where text analysis has been applied



	Bare Documents	Naturally-occurring covariates	Manually-labeled documents	Manually-built dictionaries
Latent Variable Models	Topic Models	**This Research**	Topic models with lexical constraints (like parts of speech)	
Generalized Linear Models	Document Classification or Hypotheses based on word counts			
Correlations, ratios, counts	Word Counts			Dictionary based word counts

Figure 1.3 Positioning the current research

This dissertation lies at the intersection of a sophisticated estimation method (latent variable models) and a naturally occurring (hence, relatively assumption free and less manually intensive) set of data inputs. It also serves to address an important and relatively underexplored gap in the social sciences literature on text analytic treatment of unstructured textual inputs.

Chapter 2 – Extending the LDA Approach for Improved Qualitative Analyses on Text Data

2.1 Introduction

Unstructured text typically has information content beyond what is available in standard quantitative metrics, and this information is very useful in improving our understanding of a phenomenon of interest. For example, Archak et al. (2011) extract opinions of customers about different features of products using text mining approaches. They show that these factors have a statistically significant impact on product sales. They also show that these factors improve the predictive power of models traditionally used in predicting demand.

One of the main challenges in the context of text mining is to extract the theme, or the topic from unstructured text data. The other challenge is to ascertain the importance, or the weight of the topic in the document. Many research studies used context specific approaches to address this challenge. Thus, Archak et al. (2011) identified the features of the products inherent in the review text using noun words. Further, they captured the weight of these features using the adjectives used in conjunction with the noun words. As researchers seek to expand the utilization of text mining approaches to contexts other than review texts, some challenges become apparent. First, the context may make it much harder to identify the topics in text because, for example, these topics may be more complex than product features and hence they may be harder to identify. Second, the scalability of the approach, both in terms of the scope (different contexts) as well as size (volume of text) is of great importance for the success of the text mining methodology.

Latent topic model (LTM) methods were proposed to address these challenges. Such a model assumes that there exist a number of "topic factors" – coherent semantic themes identifiable

by differential emphasis on particular words and phrases underlying any given body of text. Intuitively, given a particular topic, one would expect particular words to appear in any text document more or less frequently: Thus, in documents about consumer goods, the words ‘consumer' or 'brand' are likely to appear more often, and in documents about hi-tech goods, words like 'innovation' or 'technology' are likely to appear more often. Terms such as 'the' and 'is' that are agnostic towards a topic should appear with similar frequency in both these topics. Further, it is reasonable to expect that a document would typically contain multiple topics in different proportions. For example, in a document that is 10% about consumers and 90% about technology, it is likely that there will be about 9 times more technology related words than consumer related words. A topic model essentially captures this intuition in a mathematical framework, while making no assumptions ex ante about what the topics themselves might be. The application of a topic model facilitates both *topic discovery* in a corpus as well as the *topic composition* (or, topic proportion) of any document in that corpus.

The Latent Dirichlet Allocation (LDA) model (Blei et al. 2003) is one of most popular LTMs. The LDA model posits that all observed documents are mixtures of a finite set of topic factors, which in turn are probability distributions over word and phrase tokens. Various applications of the LDA model have appeared recently in literature (e.g., Marlin 2003; Minmo and McCallum 2007; Wei and Croft 2006; Tirunillai and Tellis 2014). Besides its use in text mining contexts, researchers have extended the application of the LDA method to non-text contexts as well. For instance, Jacobs et al. (2016) use the LDA to identify products that customers purchase together with potential applications in product recommendation systems.

Despite these efforts, the conventional LDA model suffers from certain challenges. One major challenge lies in interpreting the LDA model output to identify coherent themes, or topics.

The risk is that the method tends to overlook relatively rare tokens that have potentially topic defining properties. At the same time, high-frequency tokens, even if low in information content, are likely to influence topic interpretation disproportionately. For instance, the token ‘phone’ in a mobile device product review corpus is likely to occur frequently, and is not as useful in topic interpretation because of the many different contexts it could have occurred in. On the other hand, a token such as ‘processor’, even if it occurs infrequently, imparts more definite interpretation and meaning to a topic. To address this problem, we propose to explicitly control for a token’s occurrence probability in the corpus prior to topic interpretation. Such an approach modifies the importance of a token in terms of topic interpretation and hence verily changes the topic itself. We posit that our approach allows for superior interpretability and consequently improves quality of the downstream analysis. To illustrate how much a topic can potentially change due to the reweighting of token loadings on topics, we consider a small corpus (220 documents) of product reviews on Amazon for the Nokia Lumia smartphone from 2013. Figure 1 shows the probabilities of the most influential tokens in this corpus under both the conventional LDA and the proposed extended LDA approach. In particular, note that the influence of high-frequency tokens (‘lumia’ and ‘unlocked’) falls and that of previously low frequency but potentially informative tokens (‘seller’, ‘software’, ‘device’) rises on topic interpretation under the proposed extended LDA approach.

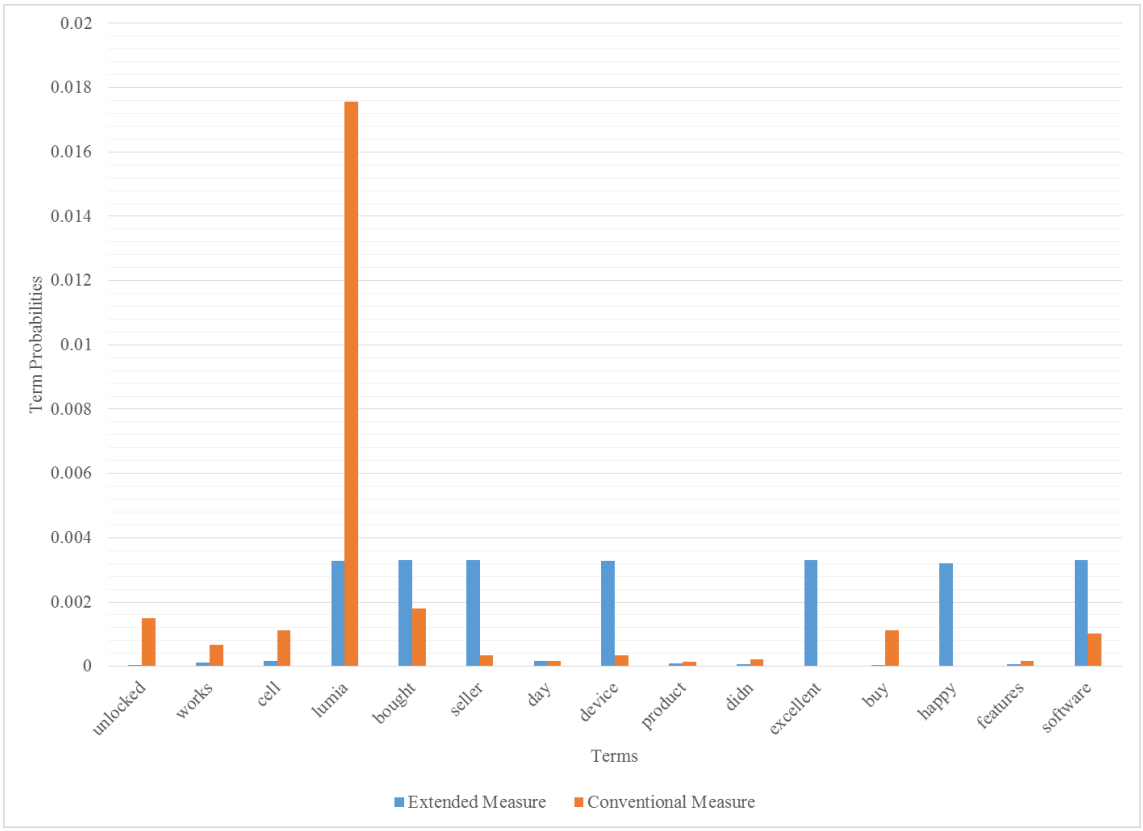


Figure 2.1. Probabilities of some influential tokens in a Mobile Review Dataset

A second limitation of the LDA approach concerns out-of-sample prediction of the topical composition of new documents. Traditionally, supervised models trained on a calibration sample's inputs and outcomes are used to predict outcomes in fresh (or prediction) samples. However, the LDA is an unsupervised model as its outputs - the topics - are latent. Hence, the LDA model must be run on any sample for which topic proportion metrics are required. In other words, it is not possible to predict an out-of-sample topic proportion unless the LDA has been run on the holdout sample as well (wherein it ceases to be 'holdout' any more). Further, even if an LDA were run on a fresh sample, the resulting topics could possibly be quite different from those in the calibration sample. This would then not only require a fresh topic interpretation exercise but would also preclude a direct, topic-by-topic comparison between the calibration and holdout samples. The

extended LDA approach we propose enables us to estimate topic proportions in new documents based solely on their constituent tokens without re-running the LDA model on the holdout corpus, thereby enabling prediction. A further benefit is that the advantages in topic interpretation discussed earlier carry-over also to the holdout sample.[1]

We need to validate and assess the extended LDA approach to establish its superiority. We do so in three ways. First, we assess the extended LDA's topic composition recovery relative to the conventional LDA approach through an extensive simulation experiment. In this simulation study, we use data from Opinosis Dataset in UCI's Machine Learning Repository to show that the extended LDA measures outperform the conventional LDA measures. Second, we analyze a product review dataset from Amazon and using human-machine and inter-human coder concordance scores, we show that topic interpretability is better in the extended LDA approach compared to the conventional LDA approach. Third, we implement our approach in the context of a well-established research theme in the IS literature, namely, the business value of IT. We empirically analyze the text of five years of 10K filings of all US public firms in the software and services industries, and obtain well-interpreted and quantified topic factors from the extended LDA approach. These topic factors indicate the strategic areas of focus for these software and service industry firms. We further incorporate these factors into an econometric model to study the impact of various inputs on firm performance. Our analysis finds that the topic factors we identified through text mining have additional explanatory power over and above the standard input factors studied in the literature. This last study is a parallel to Archak et al. (2011), who showed that information in review text had additional explanatory power over and above quantitative review

[1] We have coded all procedures for implementing the extended LDA in open-source R software. They are up on GitHub and we intend to make these procedures publicly available as interactive "shiny" R applications.

ratings. However, our study is in a different context where it is a challenge to interpret the topics due to content and context complexity.

Overall, we contribute to the literature in both methodological and substantive ways. On the methodological side, we propose and demonstrate a simple and easily applied method, based on post-processing conventional LDA output that consistently outperforms conventional LDA measures in topic interpretation and analysis. On the substantive side, we apply the proposed approach to two distinct IS research contexts, namely, thematic structure of product reviews and the Business value of IT. Thus, we provide researchers and practitioners with a set of tools to rapidly, robustly and scaleably analyze raw text corpora for enabling topic interpretation and quantification in varied contexts, and also show that incorporation of such topics is likely to improve explanatory power for within-sample analysis and have predictive power in holdout samples.

In the following section, we discuss the modeling approach and define new measures for the extended LDA approach. In Section 2.3, we utilize a simulation approach to highlight the improvement in topic interpretation due to the new measures. In Section 2.4, we show the improvement in topic interpretation using human coders with the extended LDA approach. Further, we provide an example of topic prediction using a holdout sample in this context. In Section 2.5, we analyze 10K statements of software and services firm and show that topic factors gleaned from such an analysis can improve explanatory power in established econometric models. Finally, in Section 2.6, we provide concluding remarks.

2.2 Modeling Approach

2.2.1 Background of the LDA method

Consider a text corpus **C** which is a collection of $d = 1,2..D$ distinct documents, potentially of varying length. We assume standard pre-processing routines prior to text modeling have been carried out. This entails (i) standardizing the text (by converting all text to lower case, dropping special characters, filtering out stop-words such as articles, prepositions etc.), and (ii) *tokenizing* the text into phrase-tokens (breaking down each document into 'tokens' which correspond to either single words or groups of words that occur frequently together). Now consider a document *d* with a total of N_d phrase-tokens in it. Let $w_{d,n}$ be the n^{th} token in document d. Then, we can represent the document d as a collection of tokens as: $\mathbf{w_d} = (w_{d,1}, w_{d,2}, ... w_{d,Nd})$. However, each word token in $\mathbf{w_d}$ may occur multiple times within the document. Hence, let M_d be the total number of 'unique' phrase-tokens in document *d*. Further, let $x_{d,m}$ denote the token frequency of the m^{th} of M_d tokens that occur in document *d*. Thus, $\mathbf{x_d} = (x_{d,1},...x_{d,Md})$ can be seen as a vector representation of document *d* (the vector has M_d components and each component contains corresponding token counts). By construction:

$$\sum_{i=1}^{M_d} x_{d,i} = N_d \quad \text{for} \quad d=1,2...D. \tag{1}$$

Similarly, for the corpus **C** as a whole, let $\mathbf{x_C}$ be the set of all unique phrase-token counts that occur in **C**, and let M_C be the number of unique tokens in **C**. $\mathbf{x_C}$ is given as:

$$\mathbf{x_C} = \bigcup_{d=1}^{D} \mathbf{x_d} \tag{2}$$

Suppose there exist $k = 1,2..K$ distinct, coherent and latent themes (or 'topic factors') in the corpus. Under the LDA model, these topics are essentially distributions over phrase-tokens in **C**.

Let θ_{kj} denote the probability that the j^{th} of M_C tokens belongs to (or 'loads onto') the k^{th} topic factor. Thus, by construction, $\sum_{j=1}^{M_C} \theta_{kj} = 1$. Let $\boldsymbol{\theta}_{\mathbf{k}} = \left[\theta_{k1},...,\theta_{kM_d}\right]^T$ denote the distribution of the k^{th} topic probabilities over all the phrase tokens in the sample. Aggregating $\boldsymbol{\theta}_{\mathbf{k}}$ to the entire corpus, we obtain a M_C x K dimensional matrix $\boldsymbol{\Theta}$ as $\boldsymbol{\Theta} = \left[\boldsymbol{\theta}_{\mathbf{1}},...,\boldsymbol{\theta}_{\mathbf{K}}\right]$.

The LDA model assumes that each document in the corpus is a mixture of latent topic factors. Let ω_{dk} represent the proportion of the k^{th} topic factor in the d^{th} document. Then, $\boldsymbol{\omega}_{\mathbf{d}} = \left[\omega_{d1},...,\omega_{dK}\right]$ gives the probability mass distribution of topics in document *d*. Again, by construction, $\sum_{k=1}^{K} \omega_{dk} = 1$. Aggregating $\boldsymbol{\omega}_{\mathbf{d}}$ over the entire corpus, we obtain a K x D dimensional matrix $\boldsymbol{\Omega} = \left[\boldsymbol{\omega}_{\mathbf{1}},...,\boldsymbol{\omega}_{\mathbf{D}}\right]$.

Under the LDA, we model each phrase-token occurrence vector $\mathbf{x_d}$, $d = 1,...D$ as drawn from a multinomial distribution whose arguments are a simple additive function of $\boldsymbol{\Theta}$, $\boldsymbol{\Omega}$ and N_d, thus:

$$\begin{aligned} \mathbf{x}_d &\sim MN\ \left(\boldsymbol{\theta}_{\mathbf{1}}\omega_{d1\ +\ ...\ +}\boldsymbol{\theta}_{\mathbf{k}}\omega_{dk},\ N_d\right) \\ &\sim\ MN\ \left(\boldsymbol{\Theta}\ \boldsymbol{\omega}_d,\ N_d\right). \end{aligned} \tag{3}$$

A Bayesian formulation for the model estimates the joint posterior density $P\left(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{x_C}\right)$ from the full conditional distribution on the RHS comprising the likelihood of observing data $\mathbf{X_C}$ given parameters ($\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$) denoted by $MN\left(\mathbf{x_d} \mid \boldsymbol{\Theta}\boldsymbol{\omega}_{\mathbf{d}}, N_d\right)$ times the prior distributions for the parameters denoted by $p\left(\boldsymbol{\theta}_{\mathbf{k}}\right)$ and $p\left(\boldsymbol{\omega}_{\mathbf{d}}\right)$, as follows:

$$P(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{x_C}) = \left[\prod_{d=1}^{D} MN(\mathbf{x_d} \mid \boldsymbol{\Theta}\boldsymbol{\omega_d}, N_d)\, p(\boldsymbol{\omega_d}) \right] \prod_{k=1}^{K} p(\boldsymbol{\theta_k}),$$

with priors specified as:

$$\boldsymbol{\omega_d} \sim iid \text{ Dirichlet}\left(\frac{1}{K}\right),\ d = 1,...D,$$

$$\boldsymbol{\theta_k} \sim iid \text{ Dirichlet}\left(\alpha_{k1},...,\alpha_{kM_d}\right),\ k = 1,2...K, \text{ and } \alpha_{kj} = \frac{1}{KM_d} \text{ for j=1,....}M_d. \quad (4)$$

We then estimate the model's parameter densities using a maximum a posteriori (MAP) approach (Taddy 2011) for the K topic factors. Since the LDA requires K as an input to the model, for model selection we vary K and re-estimate the model each time looking for the best model fit or the highest marginal likelihood of p ($\mathbf{x_C}$|K).

2.2.2 Extension based on the LDA Model

Consider the j^{th} of M_C phrase-tokens and the k^{th} of K topic factors. The LDA model, run on the calibration subset of the original corpus, yields θ_{kj} as the probability of j's occurrence given topic factor k's presence. In other words, θ_{kj} = Pr(token j | topic k). Since individual tokens vary enormously in their occurrence in the corpus, we control for token occurrence by normalizing θ_{kj} with the marginal probability of finding token j in the corpus, which can be written as:

$$\Pr(\text{token j}) = \frac{\sum_{d=1}^{D} x_{dj}}{\sum_{d=1}^{D}\sum_{j=1}^{M_C} x_{dj}}. \quad (6)$$

Here, $\sum_{d=1}^{D} x_{dj}$ is the total number of times j occurs in **C**, and $N_c = \sum_{d=1}^{D}\sum_{j=1}^{M_C} x_{dj}$ is the total number of occurrences for all phrase-tokens in **C**. Standardizing θ_{kj} with Pr [token j], the marginal probability of j's occurrence, yields the weight of the token for that topic. Note that if the token in relatively rare in the corpus (i.e., Pr[token j] is low), then the weight of such a token is high. Thus, the weights calculated in this way account for the rarity of the tokens and ensures that rarity does

not result in ignoring the informative value of such a token.(Netzer et al. 2012). We label this resulting measure "topic-token score", denote it by $\eta_{\text{topic.token}}$ and express it as:

$$\eta_{\text{topic.token}} = \eta_{kj} = \frac{\Pr(\text{token j, topic k})}{\Pr(\text{token j}) * \Pr(\text{ topic k})} = \frac{\theta_{kj}}{\left[\sum_{d=1}^{D} x_{dj} / \sum_{d=1}^{D}\sum_{j=1}^{M_C} x_{dj}\right]}. \quad (7)$$

$\eta_{\text{topic.token}}$ is the joint co-occurrence probability of topic and token standardized by the marginal occurrence probabilities of both topic and token. To illustrate interpretation, consider that for two distinct tokens *i* and *j*, η_{ki}/η_{kj} yields the relative magnitude of the influence of *i* in determining topic proportions compared to *j*. The $\eta_{\text{topic.token}}$ values for any topic now no longer add upto 1 (unlike the $\theta_{\text{topic.token}}$ values). However, the η's can be normalized by the column sum of the Nc x K matrix of η to obtain a probability interpretation and reach the same scale as the $\theta_{\text{topic.token}}$ values.

Using Extended LDA for Topic Prediction in New Documents

Whereas the $\eta_{\text{topic.token}}$ measure computes easily in the calibration sample (since it depends only on the calibration sample's phrase-token set), its primary significance is that it easily and scaleably extends to new, previously unseen documents as a fitted or *predictive* topic occurrence measure, $\hat{\eta}_{\text{topic.token}}$. Thus, for documents *d'*=1,...D' in the holdout sample, $\hat{\eta}_{\text{topic.token}}$ can be computed for all holdout documents, for all topics *k* = 1,...K and for all calibration corpus tokens *j'* = 1...$M_{C'}$ that occur also in the holdout corpus, thus:

$$\hat{\eta}_{kj'} = \frac{\theta_{kj} * I_{j' \in A}}{\left[\sum_{d'=1}^{D'} x_{d'j'} / \sum_{d'=1}^{D'}\sum_{j'=1}^{M_{C'}} x_{d'j'}\right]},\ k = 1,...K;\quad d' = 1,..D' \text{ and } j' = 1...M_{C'}. \quad (8)$$

Here, *A* is the set of all tokens in the calibration corpus **C** and *I* is an indicator function that takes the value of one when the holdout token *j'* also occurs in the calibration corpus, and zero otherwise.

Let $\mathbf{A}_d$ be the set of all phrase-tokens in document *d*= 1,2..D. Since documents are collections of tokens, $\eta_{topic.token}$ can be readily aggregated and summarized as a document level metric, $\xi_{topic.document}$, which is computed for topic *k* in document *d* as

$$\xi_{kd} = \sum_{j \in \mathbf{A}_d} \eta_{kj}. \tag{9}$$

The share of $\xi_{topic.document}$ metric across all topics in a document logically yields an alternative, constructive measure of topic proportions, $\kappa_{topic.document}$, which is distinct from the ω parameter in LDA output. $\kappa_{topic.document}$ for topic k and document d can be expressed as:

$$\kappa_{kd} = \frac{\xi_{kd}}{\sum_{k'=1}^{K} \xi_{k'd}}. \tag{10}$$

Further, just as for the calibration sample, the $\hat{\eta}_{topic.token}$ in holdout documents aggregates to a document level metric $\hat{\xi}_{topic.document}$ and thereafter to topic proportions in documents, namely $\hat{\kappa}_{topic.document}$. Extending calibration sample metrics to the holdout sample would only work for those phrase-tokens in the holdout corpus that are also present in the calibration corpus. Consider a document *d'* in the holdout sample. Let $\mathbf{A}_{d'}$ be the intersection of the set of all phrase-tokens in *d'* and $\mathbf{A}_C$. Then:

$$\begin{aligned} \hat{\xi}_{kd'} &= \sum_{j \in \mathbf{A}_{d'}} \hat{\eta}_{kj}, \\ \hat{\kappa}_{kd'} &= \frac{\xi_{kd'}}{\sum_{k'=1}^{K} \xi_{k'd'}}. \end{aligned} \tag{11}$$

Proposed Versus Conventional Measures

We have proposed two metrics - $\boldsymbol{\eta}_{\text{topic.token}}$ and $\boldsymbol{\kappa}_{\text{topic.document}}$ at the token and document levels of aggregation, respectively, to replace the analogous LDA parameters, $\theta_{\text{topic.token}}$ and $\boldsymbol{\omega}_{\text{topic.document,}}$ respectively. A high $\eta_{\text{topic.token}}$ would arise either due to a high $\theta_{\text{topic.token}}$, a low token occurrence probability, or both. This implies that relatively uncommon tokens, i.e. having a low token occurrence probability, coupled with a high topic membership probability ($\theta_{\text{topic.token}}$) would have higher weights in the extended LDA method, resulting in these tokens having higher influence while interpreting the topic. In the context of topic analysis in holdout samples, note that $\boldsymbol{\omega}_{\text{topic.document}}$ cannot be extrapolated to holdout sample documents *d'*, but the derivative measure $\hat{\xi}_{\text{topic.document}}$ can be viewed as the proportion of topic *k* in *d'*. To summarize, we propose to replace θ_{kj} with η_{kj} (in calibration set) and $\hat{\eta}_{\text{kj}}$ (in holdout set), and replace ω_{kd} with κ_{kd} (in calibration set) and $\hat{\kappa}_{\text{kd}}$ (in holdout set).

2.3 Simulation Experiment to Establish the Superiority of Extended LDA Metrics over Conventional LDA Metrics

To assess whether and to what extent extended LDA metrics for document level topic composition (the κ scores in equation 10) accurately recover the "true" topic composition of individual documents relative to corresponding conventional LDA measures (the ω score in equation 4), we conduct a simulation experiment. In this experiment, we choose a set of documents where we a priori know the topic composition of documents. Therefore, it is possible objectively assess the topic proportion recovery against this known “true” distribution of topics. Finally, we

also evaluate the differential accuracy in topic proportion recovery between the extended and the conventional LDA measures.

The Opinosis Dataset in UCI's Machine Learning Repository contains sentences extracted from user reviews on a given topic from among 51 different topics (Ganesan, Zhai and Han 2010). The reviews in the dataset originated from various sources - Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics). We chose three distinct and easily distinguishable topics from the 51, namely, (i) 575 reviews of rooms at the Holiday Inn, London, (ii) 333 reviews of the Netbook's battery life, and (iii) 155 reviews of the 2007 Toyota Camry's transmission. We term these three sets of documents "topical corpora" and note that they all belong to one topic each. The idea is to construct documents by mixing sentences from these topical corpora according to an ex ante known proportion. The following steps detail the construction of the simulation corpus.

Step 1. We first generate a random draw of a three-dimensional probabilities vector from a Dirichlet distribution. This vector, one for each document in the simulation corpus, is the "true" proportion of the three topics in that document.

Step 2. Because document lengths can often vary in a corpus, we randomize document length by drawing the number of sentences in each document using an exponential distribution with a reasonably large mean (30 sentences).

Step 3. We multiply document sentence length with true topic proportion to arrive at the number of sentences from each topic that should be present in each document. Then, we randomly draw this number of sentences from each topical corpus and concatenate them to construct the simulated document for which true topic proportions are ex ante known.

Step 4. We repeat this process a 100 times to generate the 100-document simulation corpus.

Step 5. Then, we run the LDA algorithm and retrieve the conventional and extended document-topic proportion measures (Ω and κ respectively).

Step 6. Next, we calculate the Euclidean distance between the "true" topic proportions for each document and the topic proportions recovered under the conventional and extended LDA measures. A smaller distance is indicative of better topic recovery.

Step 7: Finally, we sample with replacement a 100-document corpus from the Simulation corpus 10,000 times and repeat step 6 on each one. The idea is to generate a bootstrapped confidence interval around the estimated topic proportions.

Figure 2.2 shows the mean distance and 95% confidence interval for the conventional topic proportion measure (depicted as a circle) versus that for the extended model's kappa measure (depicted as a triangle) for all the 100 documents across all the iterations. Larger the mean distance, less accurate is the true topic proportion recovery. Thus, from Figure 2, we find that the extended LDA measure statistically significantly and consistently outperforms the conventional one.

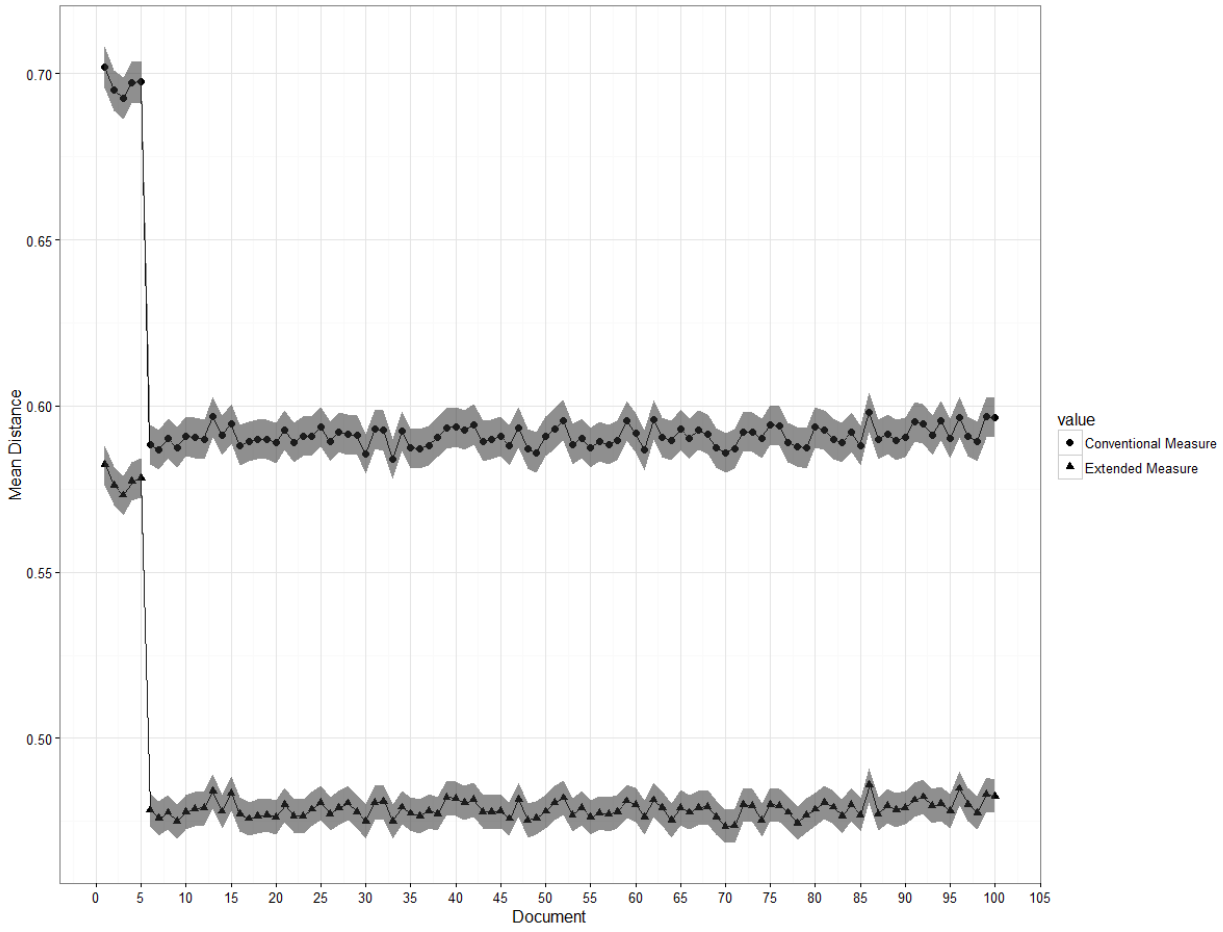


Figure 2.2. Mean Distances of LDA topic predictions from the original

2.4 Topic Interpretation by Human Coders in Extended LDA

In order to establish the efficacy of the extended LDA approach, we want to ascertain that the machine output from the extended LDA approach enables human coders to interpret the topics effectively. In order to conduct our study, we collected consumer review data from Amazon's website for a branded product launched in 2013 in the Smartphone category –the Samsung Galaxy S3. We use a relatively small data sample (about 238 reviews) to demonstrate the proposed approach, noting that larger datasets would typically strengthen the results. We split the reviews

dataset randomly into two halves – one acts as a calibration sample for estimation, validation and clustering purposes, whereas the other serves as a holdout sample for prediction purposes.

Table 2.1's Panel A shows descriptive summaries for the calibration and holdout data for the Samsung Galaxy S3. The standard deviation of the number of words per document exceeds the corresponding mean figure in both the calibration corpus and the holdout corpus demonstrating ample heterogeneity in review length. Table 2.1's Panel B shows the token-document matrix (TDM) dimensions both with and without unique tokens. Next, we ran the LDA model following Taddy (2013), on the calibration sample's TDM. We assess model fit using a Bayes factor approximation and the best fitting model's K is taken as optimal. Table 2.1's Panel C shows the log Bayes factor (or "logBF") metric used in Taddy (2013) for LTMs with K=2, 3 and 4 topics respectively. The LTM with K=2 emerges as optimal. Table 1's Panel D shows the mean, standard deviation, and range of document level $\kappa_{\text{topic.document}}$ scores for each topic for calibration data. Panel E shows the summary statistics for predicted document level $\hat{\kappa}_{\text{topic.document}}$ scores for the holdout data. Panels D and E in Table 2.1 suggest that the distribution of topic occurrence scores across calibration and holdout data is comparable. This is along expected lines since the holdout data emerged from a random split of the original corpus, and lends face validity to the $\hat{\kappa}$ measure.

Table 2.1: Finding Optimal Number of Topics – Samsung Galaxy Reviews

Group	Quantity	Value
Panel A		
Dataset Descriptive Summaries	**Calibration Dataset**	
	# Documents	120
	# words / document : mean (s.d.)	39.28 (45.42)
	Holdout Dataset	
	# Documents	100
	# words / document : mean (s.d.)	31.46 (48.39)
Panel B		
Term - Document Matrix (TDM) Size Summary	**Calibration Dataset**	
	TDM size for all terms	1813 x 120
	TDM size for repeat-use terms only	798 x 120
	Calibration Dataset	
	TDM size for all terms	1336 x 100
	TDM size for repeat-use terms only	482 x 100
Panel C		
Finding Optimal # topics	Log Bayes Factor (LogBF) for K = 2	**1415.76**
	K = 3	302.05
	K = 4	-1099.76
Panel D		
Document - Topic Scores for Callibration Data	**Topic 1**	
	Mean (s.d.)	0.529 (0.447)
	Range	[0.058, 0.953]
	Topic 2	
	Mean (s.d.)	0.471 (0.446)
	Range	[0.047, 0.942]
Panel E		
Document - Topic Scores for Holdout Data	**Topic 1**	
	Mean (s.d.)	0.477 (0.500)
	Range	[0.000, 0.999]
	Topic 2	
	Mean (s.d.)	0.522 (0.486)
	Range	[0.000, 0.999]

Note: K = 2 is optimal number of topics under the LogBF criterion

Topic Interpretation Exercise and Results

The aim is to establish that (a) there indeed exists a coherent theme that is meaningful (on the average) to humans, and (b) that the machine output captures that theme sufficiently well that it can be identified by human beings who look only at the machine output and not at the raw text.

In this vein, we assemble approximately a hundred independent human coders (graduate students) who were explained the background of the task and then exposed to machine output. This machine output, for each topic, comprised of three elements - (i) a wordcloud of phrase-tokens with highest topic membership probabilities (i.e., highest η scores), (ii) a semantic network graph of intra-document token co-occurrences, and (iii) a handful of reviews with the highest proportion for that topic (i.e., where $\hat{\kappa}_{kd'} >> \hat{\kappa}_{k'd'} \; \forall k' \neq k$). Each human coder then independently attempted to identify one coherent theme that emerges from machine output. Each coder gave the topic he/she identified (a) a brief, informative name, and (b) a description of what the topic means. Next, we collate the responses from all the coders and use judgment to classify the topics into broad 'bins' or categories. We next assess whether or not any topic or topics emerges as the dominant choice across a majority of the coders. Note that since the coders were not exposed to what the topic is or could be, and since they rely only on machine output for the topic identification exercise, we sidestep the possibility of a 'framing effect' informing the coders' analyses. We posit that the emergence of such a dominant topic from the identification task implies the existence of a clear, coherent theme that the text analysis was able to capture successfully.

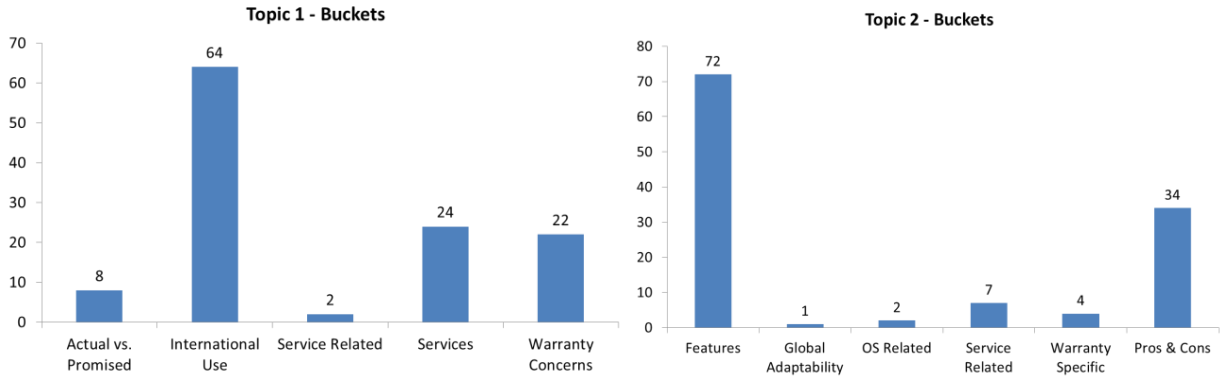


Figure 2.3. Classification of Coder Responses

Figure 2.3 shows the classification of the coders' responses. We use this classification in addition to subjectively examining the reviews to surmise that a majority of tokens and documents for Topic 1 relate to international warranty & services since they appear to correspond primarily to consumer concerns and issues regarding service warranties that come either with Samsung or with the third party resellers on Amazon. For Topic 2, a majority of tokens and documents relate to features and discussion of the issues with them. Accordingly, we label the Topics 1 and 2 as **Warranty & Service** and **Feature Specific** respectively**.** We note that a sizeable number of tokens (e.g., adjectives, verbs) are 'neutral' in that they, prima facie, may not directly inform topic interpretation. This exercise serves as external validation to our topic interpretation effort since a large number of human coders, on the average, independently arrived at similar topic interpretations from an open set of possibilities. Table 2.2 lists the topics, their labels, keywords (highest valued tokens and central graph nodes) and topic interpretation description for the Samsung S3 product review dataset.

Table 2.2. Topics for Amazon Reviews of the Samsung Galaxy S3

Topic #	Main Keywords	Suggested label	Description
1	service, great, version, unlocked	Warranty & Service	A majority of tokens and documents correspond to consumer concerns and issues regarding service warranties that come either with Samsung or with the third party resellers on Amazon
2	screen, battery, apps, iphone	Feature Specific	A majority of tokens and documents relate to features and discussion of the issues with them

Comparing Topic Interpretation - Extended versus Conventional

Further, we also want to analyze whether the topic interpretation exercise is more effective using extended LDA compared to the conventional LDA approach. In order to ascertain the relative effectiveness of the two approaches, we again recruited approximately hundred human coders from a graduate class (different from those who took part in the previous exercise) and explained to them our topic labeling and interpretation. Then we randomly split them into two groups. We showed a randomized stack of 50 phrase-tokens having the highest $\eta_{topic.token}$ scores on any one of the two latent Topics 1 and 2 to the first group. We then ask the coders to assign each phrase-token to whichever topic it most likely belongs to independently. Similarly, we show a randomized stack of 50 phrase-tokens having the highest $\theta_{topic.token}$ scores on either of the two latent topics to the second group. The idea is to compare human-machine concordance in identifying

topics membership of tokens based on two different metrics - the proposed η measure that we derive, and the conventional θ measure from LDA output. The questionnaire we use for this purpose is included in the Appendix. To assess the classification accuracy of the model output vis-a-vis human coders, we will tabulate a two-by-two table (also called the *confusion* matrix) as shown in Table 2.3. A high concordance between the model's classification and that of the human coders (i.e., high ratios in diagonal cells), indicates that the topic interpretation in terms of phrase-token propensity towards topic membership is independently validated across multiple subjects. We report hypothesized outputs in Panels A and B of Table 2.3 which show a classification concordance of 68.0% for $\eta_{topic.token}$ and 59.0% for $\theta_{topic.token}$, both of which are substantially higher than the 50% that would be obtained by random chance alone. In order to verify the $\eta_{topic.token}$ measure statistically significantly outperform the $\theta_{topic.token}$ measure on classification accuracy, we test this possibility using McNemar's test (Alpaydin 2010, pp. 501) and compare the two classification concordances. This chi-square test examines the possibility that the observed differences in classification results between the conventional and extended LDA approaches can be attributed entirely to chance. The test rejects the null hypothesis of equal classification efficacy between the proposed ($\eta_{topic.token}$) and the conventional ($\theta_{topic.token}$) measure ($\chi_{0.05,1}$ = 4.5, χ_{crit}= 5.44). Consequently, we find evidence that our derived η measure statistically significantly outperforms the corresponding conventional measure θ on topic identification and on interpretability. More generally, we contribute substantially to the broader topic mining literature by establishing the existence, subjective meaningfulness and superiority of topic interpretation for the extended LDA approach (by assessing human-machine concordance in matching word-tokens and document-content to topics).

Table 2.3. Validating Phrase-Token Classification for Samsung Reviews

Panel A		**Human Coders' Subjective Classification**	
		Topic 1	**Topic 2**
LDA Model Classification (Based on η scores)	**Topic 1 (Warranty & Service)**	**36.8%**	12.8%
	Topic 2 (Feature Specific)	19.2%	**31.2%**

Based on coders responses to top 50 phrase-tokens

Panel B		**Human Coders' Subjective Classification**	
		Topic 1	**Topic 2**
LDA Model Classification (Based on Θ scores)	**Topic 1 (Warranty & Service)**	**31.3%**	21.3%
	Topic 2 (Feature Specific)	19.7%	**27.7%**

Based on coders responses to top 50 phrase-tokens ($\chi_{0.05,1} = 5.44$)

Predicting Topic Proportions

Finally, we test the model's predictions about the latent topical structure of individual reviews (or 'documents'). We posit that the extended LDA's **κ** measure would better explain topic

proportion in documents because it is based on normalized topic-token probabilities whereas the conventional **ω** measure is not. This is a testable assumption. To test it, we sort the reviews in the descending order of their **κ** scores, randomly pick forty reviews, each of which is dominated either by Topic 1 or by Topic 2. We do the same with forty reviews picked on the basis of their high **ω** scores. Next, we split our hundred independent human coders randomly into two equivalent groups. Each coder in the first group is shown ten randomly drawn reviews from the sample of the forty highest **κ** scoring reviews. Each coder in the second group is likewise shown ten randomly chosen reviews from the sample of the top forty **ω** scoring reviews. Again, the latent topics' interpretation and labeling is explained to the respondents, who are then asked to independently and subjectively assign each review to whichever topic it most 'loaded' on. We tabulate the extent of agreement in document classification between human coders and model output in two confusion matrices - one for the first group (Table 2.4's Panel A) and one for the second (Panel B). High concordance (i.e., high ratios in diagonal cells) implies that the model's predictions of the latent topical structure of the reviews bears face and content validity in the data. Panels A and B in Table 2.4 shows a classification concordance of 60.4% and 53.2% respectively, clearly revealing that the **κ** outperforms the **ω** on document classification.

Table 2.4. Validating Document Classification for Samsung Reviews

Panel A		**Human Coders' Subjective Classification**	
		Topic 1	**Topic 2**
LDA Model Classification (Based on κ scores)	**Topic 1 (Warranty & Service)**	**28.2%**	20.4%
	Topic 2 (Feature Specific)	19.2%	**32.2%**

N = 50 responses

Panel B		**Human Coders' Subjective Classification**	
		Topic 1	**Topic 2**
LDA Model Classification (Based on Ω scores)	**Topic 1 (Warranty & Service)**	**24.0%**	26.3%
	Topic 2 (Feature Specific)	20.4%	**29.2%**

N = 52 responses

Note: *Ω scores based on combined sample for comparison*

To ensure that a particular dataset's idiosyncrasies are not driving our results, we conducted the analysis described earlier on another product reviews dataset (for Nokia Lumia 925) and found similar results. In the interest of space, we do not include these results in the paper but these are

available upon request. To summarize, in two different examples, with two different sets of topic identities, topic interpretations, and concordance among groups of human coders, we find the same qualitative result - that the derived measures η and $\boldsymbol{\kappa}$ outperform traditional LDA output θ and $\boldsymbol{\omega}$, respectively, in explanatory power (subjective topic interpretability) and in predictive power (document classification based on topic proportions).

2.5 Effect of Product-Market Strategies on Performance of Software and Services Firms

In this section, we highlight some downstream analysis based on our extended LDA approach in the context of a well-established question in the IS literature, namely, the business value of IT. We consider a real world setting wherein topic factors act as antecedents or descriptors of certain firm level outcomes. The outcome we analyze is firm performance and the topic factors we mine describe product-market choices made by the firm. Thus, our objective is to model the impact of a set of qualitative factors (topics) in product-market space on firm performance. Our aim is threefold. First, we examine whether and to what extent qualitative factors such as stated firm focus in product-market space can be measured using topic factor scores from the proposed approach as a reliable proxy. Second, we incorporate these topic factors representing firms' differential product-market focus into commonly used econometric models of dependence relations (wherein a firm performance outcome is some function of a number of determinants and antecedents). Finally, we estimate whether and to what extent the topic factors explain or describe firm performance over and above that by 'standard' antecedents. Standard antecedents include metrics such as capital expenditure (Menon et al. 2000), debt/equity ratios (Brynjolfsson and Hitt 1996), operating expenses (Mithas et al. 2012), time-varying firm characteristics, past firm

performance, etc., and control variables such as firm fixed effects, marketing expenditures (as a proxy for selling, general, and administrative expenses; see e.g., Xiong and Bharadwaj 2013) and R&D expenditures (Dutta, Narsimhan, and Rajiv 1999; McAlister, Srinivasan, and Kim 2007; Henderson et al. 2010; Brynjolfsson et al. 2002). These ideas can be summed up in a conceptual model, as shown below. Let i index firm, t time period (year) and $k = \{0,1\}$. Then:

$$\begin{pmatrix}\text{Firm}\\ \text{Performance}_{it}\end{pmatrix} = f\left(\begin{Bmatrix}\text{Past firm}\\ \text{performance}_{i,t-k}\end{Bmatrix};\begin{Bmatrix}\text{Firm}\\ \text{Characteristics}_{it}\end{Bmatrix};\begin{Bmatrix}\text{Control}\\ \text{Variables}_{i,t}\end{Bmatrix};\begin{Bmatrix}\text{Product}\\ \text{Market}\\ \text{Strategies}_{i,t-k}\end{Bmatrix}\right) + \varepsilon_{it} \qquad (12)$$

Our analysis sample comprises *all* publicly listed US firms in the Software and Services sector for a 5-year period from 2010 to 2014. We combine data from COMPUSTAT (financial variables) and the business description section (Item 1) of firms' 10-K filings (text inputs for strategy and capabilities) with the Securities and Exchange Commission (SEC). 523 firms qualify with enough observations across different data sources for the duration of the sample period. Table 5's Panel A summarizes the analysis variables from COMPUSTAT, included based on prior work in the Strategy and Finance literatures (see, e.g., Khanna and Palepu 2000; Lang, Ofek, and Stulz 1996; Yermack 1996). We use earnings per share (EPS) as our measure of firm performance for two reasons. First, EPS is widely used by analysts, investors, debtors, and others valuing firms, evaluating risks, and assessing management's performance (Dechow 1994; Dechow, Kothari, and Watts 1998; Kim and Kross 2005). Second, EPS is based on earnings which are recommended by

the Financial Accounting Standards Board (FASB 2008) guidelines which say that earnings based on accruals are superior to cash flow based measures:

> *Information about enterprise earnings based on accrual accounting generally provides a better indication of an enterprise's present and continuing ability to generate favorable cash flows than information limited to the financial effects of cash receipts and payments.*

As **X** variables, we use Cash flow from Operations ("CFO"), and number of employees ("#employees") as measures of firm and resource base size, Return on Assets ("RoA") as a measure of firm resource efficiency, Capital expenditure over sales to measure future growth opportunities ("Growth.Opp"). Additionally, as controls, we used lagged variables for marketing expenses ("Mkt.exp"), R&D expenses ("RD.exp"), and cash flow to account for the long-term effects. Tokenizing the text content of 'Item 1 - Business' in firms' 10-K filings yields term document matrices or TDMs (see Table 2.5's Panel B) of size 48,331 distinct phrase-tokens against 1,606 firm-year observations (having lost some observations when taking lagged variables).

Table 2.5: Descriptive Summary of Analysis Variables

Variable	Unit	Mean	Std Dev	Min	Max
Panel A					
Earnings Per Share (EPS)	Ratio	0.31	2.16	-27.37	25.60
Return on Assets (RoA)	Ratio	-3.61	42.76	-1050.00	316.00
Growth Opportunity (GrowthOp)	Ratio	0.04	0.48	0.00	19.00
# Employees (#Empl) [a]	'000	0.84	1.09	0.00	6.08
EBITDA [a]	'000	2.46	2.58	0.00	10.37
Cash Flow from Operations [b]	Ratio	-0.54	5.65	-153.25	101.38
Marketing Expenses (Mkt.Exp) [b]	Ratio	2.25	17.33	-2.08	355.00
R & D Expenses (RD.Exp) [b]	Ratio	0.13	0.95	-28.44	14.07
Panel B					
# of Firm Year Observations	1,606				
# of Unique Terms	48,331				

[a] *Log Value,* [b] *Scaled by Total Assets*

The LDA model applied to this TDM suggests that 5 topic factors underlie the text corpus. Upon interpretation, the topic factors appear to signify firms' product-market strategies and focus-areas. Following the proposed approach, we extracted, interpreted and labeled these 5 topic factors as (1) B2B-solution-focus, (2) online-content-focus, (3) Transaction-enabling-Capabilities, (4) Data-and-application-management-capabilities, and a fifth, "residual" topic that brings together seemingly disparate tokens that don't seem to fit well elsewhere. We propose to use this fifth, diffuse topic as the reference topic in the fixed effect portion of the econometric model (since the five topic factor scores if used together would be linearly dependent). To illustrate the interpretation process, Table 2.6 lists the highest-lift keywords, their suggested label, a reasoning for the interpretation and a representative list of well-known technology firms loading highly on the topic for the first four topics. The topic interpretation process that yields Table 2.6 relies on topic-level wordclouds and co-occurrence graphs for each topic. These details are available in the appendix.

Table 2.6. Topics, their suggested labels and interpretation

Topic #	Main Keywords	Suggested label	Description	Big firms with top loadings on this topic
1	business, solutions, management, clients, provide	B2B Solutions focus	For business clients who are global companies, to provide Management solutions such as service-offerings in sales-marketing, outsourcing, information technology, delivery, consulting etc.	Cognizant, Gartner, Pegasystems
2	internet, information, online, content, including	Online Content focus	Capabilities in Online content creation and distribution including interactive websites, games, local audiences, advertisers, consumers; especially search and commerce capabilities.	Yahoo, Trulia, AOL
3	payments, transactions, merchants, processing	Transaction enabling Capabilities	Processing and authentication capabilities for payment transactions for merchants and financial institutions in services such as accounts settlement, risk management, insurance, fees; for cash, credit, deposits	FirstData, Visa, Mastercard
4	data, software, applications, platform, product	data and application management capabilities	Enterprise-wide data application and solution capabilities for platforms and for products such as database, supply-chain organization, training etc using integrated web and software tools for applications such as user offers in real time, on-demand functionality and integration.	Oracle, RedHat, Informatica

Firms report their concurrent (or most recent period) product mix and product-market spread in their 10-K filings. Hence, the topic factors obtained from mining a given year's 10-K

statement reflect concurrent focus areas for firms in their product-market spaces. However, given that the DV, earnings-per-share is based on accruals and is backward looking, the question arises about whether a one-period lagged set of topic factors is more appropriate as an antecedent of EPS vis-a-vis a concurrent set of topic factors. We treat this as an empirical question and estimate three models which vary in the time-period of the topic factors they consider, namely, (i) concurrent topic factors only, (ii) concurrent with lagged topic factors, and (iii) lagged topic factors only. We find no significant effects for concurrent topic factors. Consequently, we report results from the best fitting model[2] (containing only one-period lagged topic scores). Let $\mathbf{TopicF}_{it}$ be the vector of K topic-specific κ scores (from Equation 10) for firm *i* in period *t*. Then, we estimate:

$$\log\left(EPS_{it}\right) = \alpha_i^{(1)} + \alpha_t^{(2)} + \alpha^{(3)} * \log\left(EPS_{i,t-1}\right) + \boldsymbol{\beta}^{(1)}\mathbf{X_{it}} + \boldsymbol{\beta}^{(2)}\mathbf{X_{it-1}} + \boldsymbol{\theta} * \log\left(\mathbf{TopicF_{i,t-1}}\right) + \varepsilon_{it} \quad (13)$$

The log-log formulation in Equation (2) is flexible, has wide use in the literature, accounts for nonlinear effects such as effect concavity (diminishing returns to Y when raising the levels of **X** variables) and allows relevant parameters to be interpreted as elasticities. We standardize the analysis variables to enable ready comparison of coefficients across variables.

Table 2.7 shows the regression results (Estimate, standard error and statistical significance) for the four relevant and interpretable topic factor terms - log concurrent **TopicF** and log (one period lagged **TopicF**). In the interest of space, we move the regression results corresponding to control variables and firm characteristics to the supplementary appendix.

[2] Results for the remaining models are available on request.

Table 2.7. Regression Results - Topic Factor Effects

Dependent variable: ***log (EPS)***		
Topic Factors	**Lag Only**	**Concurrent & Lagged**
	Est. (Std Err) **P value**	
Concurrent Topic Factors		
B2B Solutions Focus $_{t}$		0.005 (0.125) p = 0.971
Online Content Focus $_{t}$		-0.023 (0.127) p = 0.859
Data & App Management Capbilities $_{t}$		-0.018 (0.129) p = 0.891
Transaction Enabling Capabilities $_{t}$		-0.060 (0.134) p = 0.653
Lagged Topic Factors		
B2B Solutions Focus $_{t-1}$	**0.267 (0.103)** **p = 0.010*****	**0.260 (0.105)** **p = 0.014****
Online Content Focus $_{t-1}$	**0.278 (0.100)** **p = 0.006*****	**0.280 (0.102)** **p = 0.007*****
Data & App Management Capbilities $_{t-1}$	***0.197 (0.117)*** ***p = 0.093****	0.194 (0.121) p = 0.110
Transaction Enabling Capabilities $_{t-1}$	**0.189 (0.070)** **p = 0.008*****	**0.191 (0.072)** **p = 0.009*****

At least three salient points emerge from Table 2.7. First, only the lagged Topic factors are significant and not the concurrent topic factors. Thus, empirical evidence seems to suggest that a one period lag is necessary for firms' priorities and focus areas to influence firm performance outcomes in EPS terms. In other words, strategic decisions such as product-market focus which impact firm performance beyond the immediate term must allow time for the market to react to such decisions. Second, three of the four focus areas influence EPS outcomes positively and significantly at the 0.05 level (whilst the fourth does so at the 0.10 level). This suggests that, all else held constant, firms with a sharp and articulated focus in terms of product-market applications produce better outcomes relative to firms that have a diffuse product-market focus. Third, the results also reveal not just the level of statistical significance and direction (positive) for product-

market focus effects on performance outcomes, but also relative magnitude of such effects. Thus, firms having a sharp focus on *online-content* and custom *B2B-solutions* appear to do relatively better than those *Data-&-application-management* or in *transactions-enabling-&-processing*. To dig further into how much these qualitative topic factors as a whole explain performance outcomes, we conduct a variance decomposition exercise (following Silber et al. 1995). The results, shown in Figure 2.4, reveal that hitherto uncaptured cross-sectional and temporal variation in product-market foci (which we now are able to measure as topic factor scores) contributes to a little over 4% of explained variance in firm performance outcomes. This estimate is after accounting for firm-specific fixed effects and other control variables.

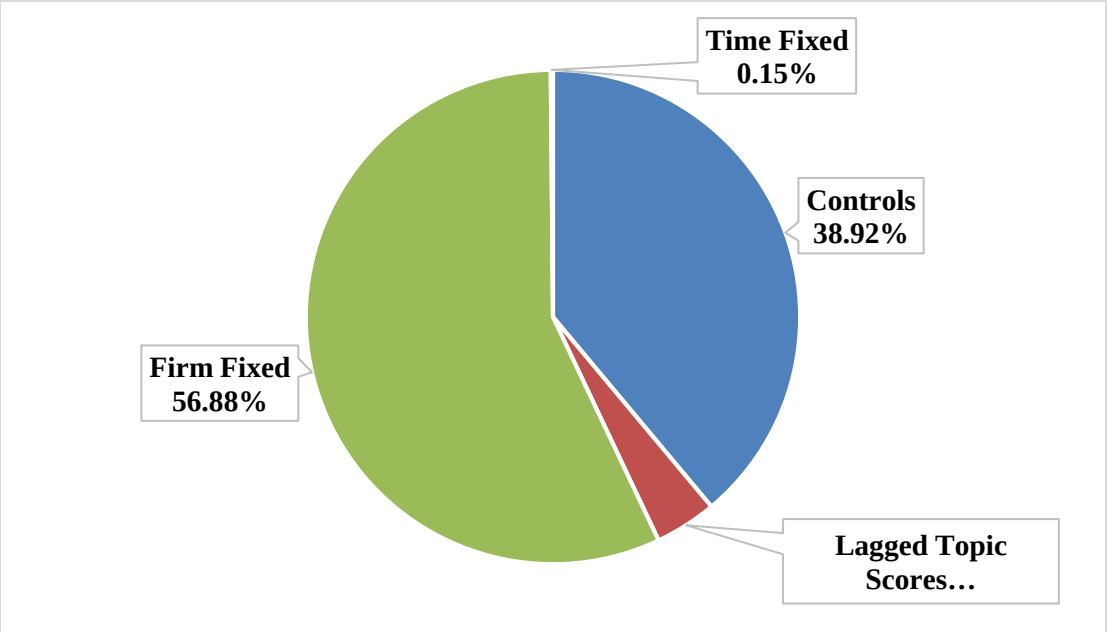


Figure 2.4. Variance Decomposition

To summarize, we present a brief and illustrative application of the proposed topic identification and extraction method to a well-known question in IS research. This application highlights that downstream analysis using econometric models in varied contexts is possible. More generally, through this study we demonstrate the possibility of adding a host of new independent variables that capture previously hard-to-measure qualitative considerations and incorporate them into extant econometric models, in complex and widely varied contexts, using a relatively quick and scalable approach enabled through extended LDA.

2.6 Concluding Remarks

2.6.1 Summary

We present researchers and practitioners with a powerful approach to systematically glean insights from unstructured text found in data sources readily accessible to firms. Our approach builds upon the popular LDA model for topic mining. We empirically demonstrate the approach using two distinct datasets - Amazon Product reviews and the 10-K statements of Technology firms. We present both methodological and substantive contributions to research.

On the methodological side, whereas several variations and extensions of the LDA have been developed over the years, our approach post-processes LDA output to yield derivative measures which achieve: (a) superior topic interpretation compared to conventional LDA output (in particular, we emphasize occurrence normalized token probabilities $\eta_{topic.token}$ and the use of token co-occurrence graphs to aid topic interpretation); (b) superior document classification based on topical structure; (the derivative $\boldsymbol{\kappa}$ measure consistently and significantly outperforms it's

conventional counterpart **ω** in document classification); (c) predictive power in holdout samples for previously un-analyzed documents that is otherwise unavailable in conventional LDA output; and (d) the incorporation of semantic themes as metric variables into standard econometric model specifications. In addition, our proposed approach relies entirely upon open source platforms and tools, and thereby presents researchers and practitioners a means to readily adopt and transfer the approach across teams, geographies and organizations without the hassle of bringing in proprietary components and their licensing requirements.

On the substantive side, we apply the extended LDA approach to a very relevant question in IS literature, namely, estimating the business value of IT. We discover five distinct, coherent themes indicative of product-market strategies in technology companies' 10-K filings of which four are meaningful. We measure and interpret these topics and use them as antecedents in an econometric model of enterprise performance. We find positive and statistically significant effects of the product-market strategies on firm value. Our approach also offers researchers and practitioners a substantive aid to build models and test hypotheses of interest that involve mining open-ended text. Since the approach yields metric output, enabling all manner of downstream statistical analyses. An additional substantive contribution is investigating and establishing the subjective existence and validity of topic interpretation (by assessing human-machine concordance in matching word-tokens to topics).

2.6.2 Limitations and Future Research

The research has limitations which future research could address. One limitation endemic to text mining methods in general is that the choices made during pre-processing impact the size and composition of the Term Document matrices and thereby can potentially influence topic discovery. Some of these choices include (a) pre-setting frequency settings for whether or not to

include either unique tokens (which occur only once in the corpus) or extremely low frequency tokens, (b) the choice of 'n' in mining for n-grams (e.g., 'service' is a unigram word-token whereas 'service provider' as a phrase-token is a bigram and so on), (c) the quality of stopword lists used (e.g., stopwords such as articles or prepositions can be filtered out of the corpus prior to tokenization), and so on. These modeling choices are often situational and contextual. Though we performed standard robustness checks, the large number of possible modeling choices can be a challenge. A second limitation concerns the prediction of topic proportions in holdout samples. Here, the caveat is that any tokens in the holdout sample that never appeared in the calibration sample will necessarily drop out of the analysis. A third limitation is that we used only two datasets. A greater number of text corpora from multiple categories in B2C and industries or sectors in B2B would help understand how topic discovery and interpretation varies from one context to another within the B2C and B2B rubrics. A fourth limitation (and related future research avenues) pertains specifically to the BVIT application in Study 2. The application of the extended LDA approach for capturing firm capabilities in the technology space and utilizing them in econometric models throws open a number of questions and implications. For instance, one may ask what theoretical basis exist for certain latent topic factors to impact firm performance. Are there empirical regularities, patterns of interest etc. in the distribution of latent topic factors (corresponding to product-market space location choices and capabilities)? We leave these and other such questions in different application contexts to further research.

Chapter 3 - Brand Marketing Effectiveness, Brand Positioning Gaps and Brand Outcomes: An Empirical Investigation

3.1 Introduction

A central question in branding research concerns the marketing function's effectiveness in impacting brand outcomes. Recent work by Petersen et al. (2017) suggests that there is a renewed interest in studying customer mindset metrics but there is a lack of understanding of how and to what extent do these metrics translate into profitability. Also, the role of these metrics to explain shareholder value has been underexplored (Colicev et al., 2018). One of the challenges faced by marketing is to correlate brand awareness, brand loyalty and sales revenue to evaluate the effectiveness of its actions (Hanssens & Pauwels, 2016). In this chapter I propose to measure the marketing effectiveness from unstructured data to better understand its implications for firms and brands. In order to do so, I propose the "positioning gap" as a measure of marketing effectiveness.

I define the "positioning gap" as the difference in intended and actual positioning of firms. To arrive at this measure, I include all types of discussion captured in unstructured format-consumer opinions as captured in consumer reviews and blogs, intra firm discussions and, public communications by firms and regulatory institutions. I combine econometric methods and natural language technologies to examine the influences on and of "perception gap". I use the extension to a popular latent topic modeling algorithm (proposed in chapter 2) to extract insights about "intended positioning" from public communication by brands and "actual positioning" from consumer opinions, and newspaper and magazine articles. As depicted in Figure 3.1, I use the insights thus generated to (i) arrive at a measure of the distance between the two positionings – "positioning gap", and (ii) classify the positioning mismatch to uncover insights into reducing the

gap. I also use the gap along with standard co-variates and controls to explain how much it contributes to firm performance.

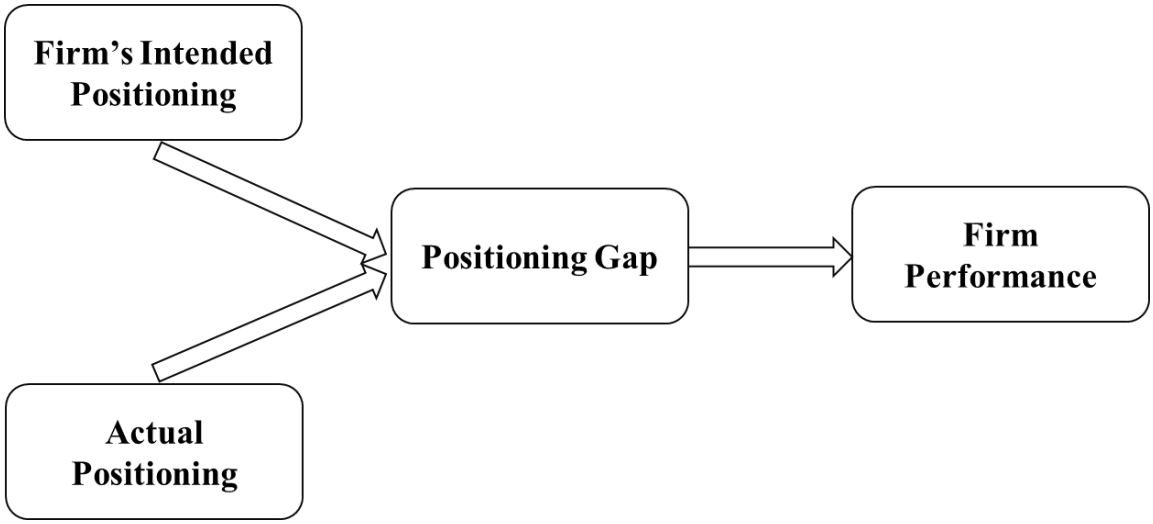


Figure 3.1: Framework

3.2 Background and Literature Review

Branding research is concerned with the effectiveness of marketing in impacting outcomes. Brand managers are concerned about improving performance including sales, rank, market share, and brand equity. The macroeconomic conditions, product category and industry conditions cannot be influenced by the brand managers, however, they can influence the brands' own characteristics. In other words, this relationship can be expressed as:

$$\begin{pmatrix}\text{Brand}\\ \text{Outcomes}\end{pmatrix} = f\left(\left\{\begin{matrix}\text{Brand}\\ \text{characteristics}\end{matrix}\right\};\left\{\begin{matrix}\text{Control}\\ \text{variables}\end{matrix}\right\}\right) + \varepsilon \quad \ldots\ldots (1)$$

The brand characteristics and be further categorized as subjective characteristics such as the brands' age, diversification, and past performance, and objective characteristics such as brand trust, perceived quality, and brand persona. Thus, equation 1 can be rewritten as:

$$\begin{pmatrix}\text{Brand}\\ \text{Outcomes}\end{pmatrix} = f\left(\left[\begin{Bmatrix}\text{Objective Brand}\\ \text{characteristics}\end{Bmatrix};\begin{Bmatrix}\text{Subjective Brand}\\ \text{characteristics}\end{Bmatrix}\right];\begin{Bmatrix}\text{Control}\\ \text{variables}\end{Bmatrix}\right)+\varepsilon \quad \ldots. (2)$$

A brand manager has limited control over the objective characteristics, but can influence the subjective characteristics and hence the brand outcomes. In order to include the subjective brand characteristics, the brand manages targets his communication and messaging to achieve a certain image (positioning) among her customers. We can thus argue that the net marketing effect on brand characteristics is the difference between the intended and achieved positioning of the brand as a result of the brand managers' communication strategies.

$$\sum\begin{pmatrix}\text{Marketing Effects on}\\ \text{Brand Characteristics}\end{pmatrix} = g\begin{pmatrix}\text{Intended vs Achieved}\\ \text{Brand Positioning Gap}\end{pmatrix} \quad \ldots (3)$$

3.2.1 Positioning and Positioning Gap

Positioning refers to a brand's location relative to competing brands in the minds of customers, who evaluate brands along various attribute dimensions (Doyle 1975). The notion of positioning implies a space where the points are the brands and the dimensions of the space are attributes which are salient discriminators for consumers. In general, such a space will be multidimensional, and brands will be judged on more than one attribute. Figure 3.2 depicts one such space which is also referred as a positioning map. Such positioning maps are often represented in the two-dimensional space with the X and Y axes depicting the two of the most preferred attributes for the brands under consideration.

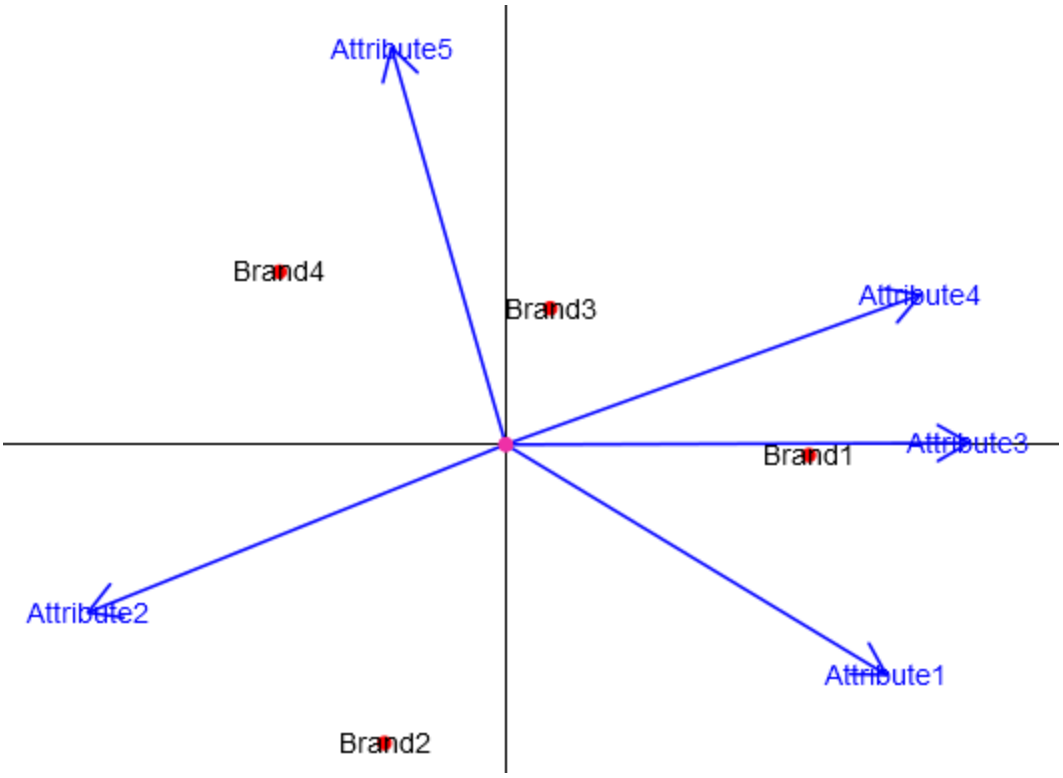


Figure 3.2 Positioning Map

A spatial representation, as in Figure 3.2, has a number of advantages (Doyle 1975) including:

i) Evaluating brands along these attributes and the relative strengths and weakness can be estimated

ii) Suggesting preferred positions for brands in the space and evaluating opportunities.

Positioning Gap

We argue that firm's intended positioning - such as management's perceptions, priorities, plans, goals, constraints, strategies - are captured to a substantial extent in a firm's own vocabulary. These considerations influence managerial decision making, which in turn drives firm behavior and thereby, firm outcomes. We leverage text analytic methods to codify and measure managerial focus in the Marketing domain. An effective measure of the "positioning" that is salient in managers' mind is to consider the "positioning" as perceived by the consumers (end users, expert

reviewers, critiques, or media articles). Leveraging the same text analytic methods using to codify and measure the managerial focus, we determine a measure for the consumers' focus. Having codified the two sources of textual intent, we propose a measure of marketing effectiveness as the difference in the two calculated measures. We then hypothesize that smaller this calculated measure, the higher is the marketing effectiveness. The higher is the marketing effectiveness the higher will be its effect on firm's value while controlling for standard co-variates.

$$Positioning\ Gap_i = f(Intended\ Positioning_i, Actual\ Positioning_i) \ldots\ldots (3)$$

Thus a *positioning gap* is the difference between the brand's intended versus its achieved position in customers' minds. Measuring the gap using text data alone helps overcome the challenges that arise when gathering similar primary data through surveys wherein issues such as cost, complexity, comparability across categories and domains, and scalability arise. Our proposed approach not only uses secondary data but is also scalable to (i) measure both the intended and the achieved positioning, (ii) for a large number of brands, (iii) having potentially differing attribute dimensions, and (iv) even across categories and domains.

3.2.2 Literature Review

Various researchers have studied the effectiveness of marketing using a collection of measures. These measures are either originating from the brand which we label as *Brand Generated Content (BG)*, or originate from other sources such as customers and print media which we label as *Organically Generated Content (OG)*. Table 3.1 summarizes the various studies relevant to this chapter. Researchers in the past have classified the BG content as those originating from traditional advertising (Bruce, Foutz and Kolsarici 2012; de Vries, Gensler & Leeflang 2017) or those originating from brands' social media (Kumar et al. 2016; Goh, Heng & Lin 2013; Pauwels et al.

2016; Colicev et al. 2018). These same researchers have used users' social media content as a source for the OG content. In this chapter, in addition to these existing sources, we propose the use of traditional media and press as a source of OG content. Like has been used by researchers in the past, we use unstructured textual data available publicly on the internet.

		Organically Generated Content (OG Content)		
		None	**Users' Social Media**	**Traditional Media and Press**
Brand Generated Content (BG Content)	**Traditional Advertising**		Bruce, Foutz and Kolsarici (2012), de Vries, Gensler & Leeflang (2017), **This Chapter**	**This Chapter**
	Brand Social Media	Kumar et al. (2016)	Goh, Heng & Lin (2013), Pauwels et al. (2016), Colicev et al. (2018), **This Chapter**	**This Chapter**
	None		Tirunillai & Tellis (2012), Kumar et al. (2013)	

Table 3.1 Literature Summary

3.3 Data

We aggregate our data from various sources. First, we use the top 250 global brands as provided by brandirectory.com as our sample. We use the firm performance and other financial variables from Euromonitor. We then use the brand name as our search term on Google search to identify the top 50 trending sources of data for the brands. In the merging process, we lose a few

brands and are left with 197 brands in the final sample. Having identified the sources, we classify manually them as BG and OG to create our two sources of textual data. Table 3.2 below provides summary statistics for all the textual data. Even after losing a few brands, we are left with a large enough text corpus to arrive at our measure marketing effectiveness.

Total Brands	197	
	BG Content	**OG Content**
# words / document (average)	759	2,130
# words / document (std dev)	1,298	14,763
# words / document (max)	54,443	741,255
Sentiment Score (ave)	0.279	0.172
Sentiment Score (std dev)	0.273	0.269
Sentiment Score (min)	-0.723	-0.985
Sentiment Score (max)	0.996	0.999
# documents / brand (Ave)	25	38
# documents / brand (Min)	1	1
# documents / brand (Max)	84	106
# documents / brand (Std Dev)	14	17
# documents	4,858	7,507
Unfiltered TDM Size	417,698 x 12,365	

Table 3.2 Summary Statistics

3.4 Modelling Approach, Results and Analysis

Deriving Metric Measures from the Textual Data

The first step in the process of arriving at our proposed measure is the calculate the positioning gap. To arrive at this, we need to first estimate the positioning map for the BG content and OG content for each of the brands. For each of the brands, stack together the BG and OG content and then use the method proposed in chapter 2 to arrive at the various dimensions that are latent to the text data collected. The intended positioning and actual positioning are a vector of $n \times k$ topic scores derived from the textual data using LDA based topic mining, where n is the total number of textual data observations (BG and OG) stacked together and k is the number of latent topics or themes in the underlying text. Using this information, we then derive the following measures:

1. Size of the BG cloud
2. Size of the OG cloud
3. Centroids of BG and OG clouds
    a. The distance between the two centroids gives us the measure of the gap. This is what we define as the *positioning gap* – defined as the Euclidian distance between the BG and OG centroids.

In the next section I provide a motivation for the BG and OG cloud along with other text-based measures that are used in further analysis.

Other Measures of Textual Diversity

Since we are using the unstructured textual data available on the internet as a proxy for the open ended survey responses to arrive at the brands' intended and achieved positioning, it is also

important to use the data to extract other relevant measures of textual diversity. Research in linguistics, especially McNamara, Crossley, and McCarthy (2010), as shown in Figure 3.3 argue that text content with a more complex syntax, with greater lexical diversity and less frequent words is not only more difficult to understand but also implies the use of more sophisticated language.

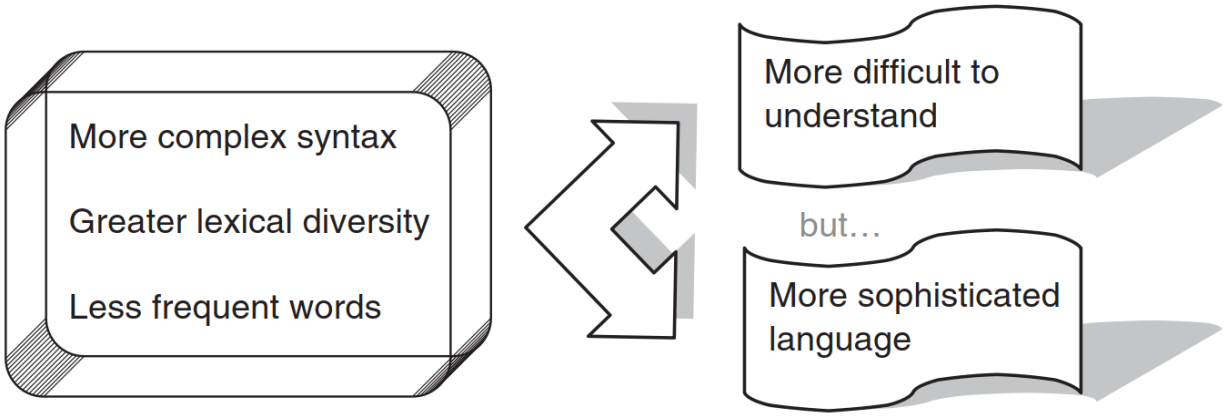


Figure 3.3 McNamara, Crossley, and McCarthy (2010) on Complexity of Textual Data

They argue that the complexity of the textual content can be understood better using the four components:

i. Syntactic Complexity - set of rules, principles, and processes that govern the structure of sentence, usually also include the order in which the words occur

ii. Lexical Diversity - ratio of different unique word stems (types) to the total number of words (tokens)

iii. Word Frequency – total number of words (tokens)

iv. Sentiments - aims to determine the overall contextual polarity or emotional reaction to a document, interaction, or event

In this chapter, I use the number of themes identified by the topic modelling exercise as a measure of the syntactic complexity, the lexical diversity scores[3]as a measure of the lexical diversity, word counts and sentiment scores as the measures to better understand the structure of the textual content.

Explaining the Brand Value

Having defined and extracted the measures of interest from the textual data, I then proceed to show that the measures help in explaining the brand values. For this, I propose the use of the following equation:

$$BV_t = \beta^{(1)} BV_{t-1} + \beta^{(2)} TypeOfGood + \beta^{(3)} Product + \beta^{(4)} Age + \beta^{(5)} Newness + \beta^{(6)}\ PositioningGap + \beta^{(7)}\ SizeofBGCloud + \beta^{(8)} SizeofOGCloud + \varepsilon \quad \ldots\ldots (4)$$

Where,

BV is the Brand Value as reported by branddirectory.com,

TypeOfGood is a variable classifying the brand as an experience good (0) or Search Good (1),

Product is a categorical variable classifying the brand as a Product (1), Service (2) or Mix (0),

Age is the age of the brand,

Newness is the newness of the brand relative to its category,

[3] Please see the appendix for a list of all available Lexical Diversity Measures and their definitions

Positioning Gap is the measure of marketing effectiveness calculated from textual data,

SizeOfBGCloud is the dispersion of the BG content, and

SizeOfOGCloud is the dispersion of the OG content.

The results as show in table 3.3 suggest the positioning gap has an inverse relationship with the brand's value. The results help ascertain the fact that higher is the brand's marketing effectiveness (i.e. small positioning gap), the better is the brand's performance (i.e. brand value).

Coefficients	**Estimate**	**StdErr**
Intercept	-1.6060	1.5220
Brand Value $_{t-1}$ #	**1.1550**	**0.1169 *** **
Type of Good (Expereince = 0, Search = 1)	0.1742	0.1182
Product or Service (Product = 1, Service = 2, Mix = 0)		
Product	0.1640	0.2425
Service	0.3802	0.2576
Age #	0.0027	0.0665
Newness Realtive to Category	0.0005	0.0011
Positioning Gap	**-2.7020**	**1.4700 ***
Size of BG Cloud	0.1288	0.3442
Size of OG Cloud	-0.3556	0.3058
N	197	
# Log	* p<.05 ** p<.01 *** p<.001	

Table 3.3 Explaining the Brand Value

Explaining the Positioning Gap

Having ascertained that the positioning is a valid and significant measure of brand's marketing effectiveness, I next proceed to understand the drivers of the positioning gap. This section will help arrive at actionable insights for managers. For this, I include the textual

diversity measures motivated earlier to arrive at the complexity of the textual data and explain the positioning gap. Table 3.4 summarizes the results of the analysis. The variables are grouped into three sub-groups:

1. Variables related to both BG and OG content
    a. **# of Distinct Themes**: I find that the higher the number of latent themes in the text data, the smaller is the positioning gap. The higher number of latent themes is a proxy for complex but water-tight positioning of the brand by the managers. This is reflective of the manager understanding her customers better and in turn achieving a better position.
2. Variables Related to BG Content
    a. **Word Count**: The word count of the BG content has a quadratic relationship with the positioning gap. The relationship suggests that while brands can be verbose when defining their position, after a certain threshold, it negatively impacts the effectiveness. This goes well with the earlier result about water-tight positioning and the use of fewer words but a complex syntax.
    b. **Sentiments**: The results suggest that sentiments play a role in reducing the gap.
3. Variables related to OG Content
    a. **Lexical Diversity**: The measure is reflective of the fact that when customers uses a diverse set of tokens to express their views on the brand, it helps narrow the positioning gap thereby improving the marketing effectiveness.

To summarize, we show that the positioning gap as defined is an alternative measure of marketing effectiveness. This can be easily scaled to include more brands without incurring any additional costs to conduct additional interviews with managers or surveys from customers. We

also explain actional insights for managers in order to help them reduce the gap between the intended and actual positioning and hence improving the marketing effectiveness.

Coefficients	Estimate	StdErr	
Intercept	0.1242	0.0067	
Variables Related to both BG and OG Content			
No. of Distinct Themes	**-0.2419**	**0.1199**	*****
No. of Distinct Themes^2	-0.0881	0.1095	
Message Sentiment			
Variation	-15.1457	11.9289	
Variation ^{2}	***-7.3301***	***4.2847***	**.**
Average	-0.8238	2.0578	
Variables Related to BG Content			
Size of Cloud	***-0.3262***	***0.1925***	**.**
Size of Cloud^2	0.1658	0.1269	
Lexical Diversity	-0.1871	0.1271	
Lexical Diversity^2	-0.1115	0.1041	
Word Count	**-0.3523**	**0.1216**	******
Word Count^2	**0.5660**	**0.1229**	*******
Message Sentiment			
Variation	**-0.9667**	**0.3625**	******
Variation^2	**0.6546**	**0.1904**	*******
Average	**0.6822**	**0.3257**	*****
Variables Related to OG Content			
Size of Cloud	8.5709	9.1712	
Size of Cloud^2	1.5445	1.6395	
Lexical Diversity	**-0.4139**	**0.1248**	******
Lexical Diversity^2	0.1209	0.1024	
Word Count	-0.0272	0.1396	
Word Count^2	-0.0021	0.1159	
Message Sentiment			
Variation	10.3592	8.7652	
Variation ^{2}	***5.2508***	***3.0312***	**.**
Average	3.4631	3.2297	

^{2} Non - linear (quadartic) component * p<.05 ** p<.01 *** p<.001

Table 3.4 Explaining the Positioning Gap

3.5 Limitations and Future Research

This research has certain limitations which future research can address. One of the biggest challenges in building this dataset has been to collect the textual data from publicly available sources for each of the top brands. While I have been able to show that using text scraping methods, such data can be assimilated from publicly available sources from the internet, the analysis is presented for one time period, i.e. is a cross-sectional dataset. Future research could look at collecting historical text data and also train computer algorithms to collect data for the future in order to create a rich dataset which will also help explain the time related dynamics in this relationship.

Chapter 4 – The Dynamics of Product – Market Choices in the Technology Sector

4.1 Introduction

The determinants of firm performance have elicited a lot of research attention, particularly in the economics and management literatures. A widely accepted perspective is that firm performance is some function of factors internal to the firm (e.g., firm demographics, organizational structure, capital structure etc.), of factors external to the firm such as environmental variables (e.g., industry structure and characteristics, regulatory environment, macroeconomic conditions etc.), and of firm strategy, which connects and reconciles these internal and external factors (e.g., Grant 2008). For this chapter, we limit our scope of strategy only to that part which is a direct result of firm's marketing strategy. In the rest of the chapter, all references to firm's strategy are limited to the firm's marketing strategy. In conceptual terms, we could say:

$$\begin{pmatrix}\text{Firm i's}\\ \text{Performance}\end{pmatrix} = f\left(\begin{Bmatrix}\text{Internal}\\ \text{Factors of i}\end{Bmatrix}\begin{Bmatrix}\text{External Environment}\\ \text{of firm i}\end{Bmatrix}\begin{Bmatrix}\text{Firm i's}\\ \text{Strategy}\end{Bmatrix}\right). \quad (1)$$

However, firm strategy is an intangible quantity whose measurement, modeling and analysis have proven to be a challenge. Managers, when asked to describe firm strategy, do so in words. Most external communication by firms, used for informing various stakeholders the firms' strategy, is textual in form. Firms' compliance filings, made at the behest of the government and regulators, often contain strategic content of interest and are again primarily textual in form. Hence, under the assumption that what firms are saying about their strategy is significantly correlated with an underlying, latent strategy construct, we could say in conceptual terms:

$$\begin{pmatrix}\text{Firm i's}\\ \text{Performance}\end{pmatrix} = f\left(\begin{Bmatrix}\text{Internal}\\ \text{Factors of i}\end{Bmatrix}\begin{Bmatrix}\text{External Environment}\\ \text{of firm i}\end{Bmatrix}, \begin{Bmatrix}\text{Firm i's Strategy in a}\\ \text{descriptive, textual form}\end{Bmatrix}\right). \quad (2)$$

Following equation (2), we develop a robust and general approach to measure qualitatively-described variables of interest (firm strategy, in the current context), model their effects on a set of units of interest (firms, in our case), and invoke well-established estimation and statistical inference techniques to assess the marginal effect of a text-based quantity on outcomes of interest (firm performance metrics). Next, we describe the challenges typically faced in strategy measurement, and how our approach addresses the same.

4.1.1 Measuring firm strategy

A number of approaches have been employed to measure or account for strategy in formal analysis. One approach has been to include metric variables such as those reported by the firm in financial statements and annual reports as proxies for focal elements of Strategy. For instance, to measure firms' strategic innovation emphasis or orientation, studies have used various proxy variables such as internal R&D spend (e.g., Cassiman and Veugelers 2006; Bronzini and Piselli 2016), the number of patents filed (e.g., Schilling and Phelps 2007; Griliches 1990; Helmers and Rogers 2011), new product introductions (e.g., Basberg 1987), etc. Another approach has been to conceptualize particular aspects of strategy followed by theory-building and testing using either primary data from managers or secondary data about firms. For instance, to measure the impact of mission and vision statements (which are purely textual information) on firm performance, various studies have used theoretical conceptualization (e.g., Baum et al. 1998), coding the statements and

classifying into pre-categorized bins (e.g., Collins and Porras 1991), interviews with entrepreneurs (e.g., Filion 1991) or with CXOs (e.g. Larwood et al. 1995), content analysis (developed by Winter and McClelland 1978), scoring (e.g., Kirkpatrick et al. 2002), etc. This approach brings with it challenges of cost, complexity, and scalability, among other things. We take an alternative, general, and scalable approach. We mine for and directly transform strategic content in relevant textual sources into usable econometric variables. The elements of strategy focal to our study are most readily available in the form of descriptive text in the firm's 10-K filings with the SEC.

First, we put together a repository of potential data sources for strategic text content such as firm filings with the SEC, annual reports, transcripts of analyst meets, etc., that firms put out in the public domain to comply with legal and regulatory requirements. Next, following recent advances in text analytic procedures (e.g., Hoberg and Philips 2010), we "tokenize" the text corpus, i.e., transform each firm's text description of strategy into a vector of "Strategy phrase-tokens". This vector of word tokens constitutes a large-dimensional set of metric variables corresponding to the text description of a firm's strategy. We propose that each firm's text vector denotes its "location co-ordinates" in a latent text space (latent strategy space in this case), and that every firm in the analysis sample can be simultaneously positioned in this latent strategy space. Using the "extended LDA" method described in chapter 2, we dimension reduce these vectors to coherent themes to give us the "product-market choices". Here, we make the implicit assumption that similarities in the vocabulary used to describe firms' strategic choices (both intended and realized) correlate with similarities in firm strategy. We thus have each firm's location relative to every other firm in the analysis sample in latent product-market choices space. Although applied in the context of firm strategy in latent strategy spaces, the proposed approach is general and readily extends to any textual data in relevant latent space for any units of analysis.

Our goal is to estimate the marginal effects of firm strategy on firm performance (henceforth, FP) while controlling for the FP effects of other factors and, thereafter, perform analysis, interpretation, and inference on the results. This requires the development and deployment of a formal model that connects firm strategy to FP within the constraints outlined thus far. We develop a variant of a linear model to incorporate text attributes into formal analysis. We use relative firm locations in latent strategy space as the product-market choice (henceforth, PM choice) and thereby enable estimation, inference, and interpretation of inter-firm strategy effects on FP. We present the description of the model output in the Model section. The data comes from publicly traded firms in the Technology sector in the US. The PM choice is mined from firms' 10-K filings with the SEC. The firm performance data comes from Compustat. We find that accounting for PM choices uncovers significant effects of firm characteristics on firm performance.

4.2 Positioning the Research

The determinants of firm performance have elicited a lot of research attention, particularly in the economics and management literatures. In this section, I position my research among the past research and also highlight its contributions to the extant literature. Table 4.1 focuses on past research which has used one or more measures based on text analytics and has used stock market or accounting based measures. For example, Nath and Mahajan (2008), and Germann, Ebbes, and Grewal (2105) find contradictory results for CMO presence on FP but do not use any text analytics based measures. A recent working paper, proposes the marketing mindset metric (Worm, et al.) which is the most advanced measure derived using textual data to arrive at a metric measurement of marketing intent in the top management to study its effects on FP. In this chapter, I propose the use of textual data to derive the latent PM choices. I then use these PM choices to determine firms

that have chosen the popular choices and those that have made niche choices. I study these choices for a ten year period between 2005 – 2014 in order to classify firms into groups based on their decision to switch between choices or maintain the same strategy during the time period. A propensity score based matching exercise helps estimate the causal effect of such a decision by the firms on their FP.

		Text Analytics		
		No	**Basic**	**Advanced**
Stock Market OR Accounting Based Measures	**No**	**NAICS Codes based** Wang & Zajac (2007) **Market Complements** Mitshuhashi & Greve (2009)	**CEO's Attention to Future** Yadav, Prabhu & Chandy (2007) **Similarity in Business Descriptions** Hoberg & Philips (2010)	**Business Proximity** Shi, Lee & Whinston (2016)
	Yes	**CMO Presence** Nath and Mahajan (2008); Germann, Ebbes & Grewal (2015) **Marketing Department Power** Feng, Morgan & Rego (2015)	**Marketing Mindset** Worm, Bhardwaj, Shen, Srivastava (Working Paper)	**Product – Market Choices** This Paper

Table 4.1 Reviewing the Relevant Literature

4.3 Solution Approach

In this chapter, I utilize the textual data in the form of 10-K filings of the firms with the Securities and Exchange Commission (SEC) to mine for the product-market choices by the senior management. While some might contend that the use of annual reports is less than optimal, they still are the best source of information on senior management's cognitions available to researchers (Abrahamson and Hambrick 1997). One potential threat to the validity of information in annual reports could arise if their content were shaped by the communications agencies involved rather than by the management. However, Abrahamson and Hambrick (1997) find strong evidence that senior executives are highly involved in writing annual reports. In line with this finding, a large number of studies confirm the nomological validity of cognitive measures extracted from annual reports (e.g., Cho and Hambrick 2006; Daly, Pouder, and Kabanov 2004; Eggers and Kaplan 2009; Osborne, Stubbart, and Ramaprasad 2001; Yadav, Prabhu, and Chandy 2007). Another potential threat could lie in executives' attempts at impression management through annual reports. Multiple studies have however countered this criticism, showing that patterns of attribution in annual reports reflect managers' actual cognitions (Clapham and Schwenk 1991; Fiol 1995). The supporting evidence from these studies justifies our choice of 10-K reports to mine for the product – market choices.

4.3.1 Locating Firms in the Product – Market Space

I first build a product-market choices space for the entire US technology sector by text mining for latent structure in the Business description section of firms' annual 10-K compliance filings with the SEC. I use the method proposed in chapter 2 for mining the unstructured textual

data. As in chapter 3, I stack all the textual data to mine for the themes which are the product – market choices made by the managers in the time period under consideration. Following an argument similar to that in chapter 3, I then treat the product-market choices thus arrived as the dimensions of a multi-dimensional product-market choices space. Next, I locate each firm in each year in this product-market choices space relative to every other firms' location. This then helps identify changes or movements in firm positions in product-market space over time. It also helps identify regions of dense (popular) versus sparse (niche) competition resulting from firms' product-market choices. Figure 4.1 shows a sample product-market choices space. It considers that firms in the technology sector are faced with 3 choice – (i) cloud computing, (ii) core computing, and (iii) hybrid computing. The dark triangles represent cloud computing, the dark circles represent core computing whereas the lighter shapes represent hybrid computing. Firms that are in the dense regions are the ones which are making the popular product-market choices whereas those that are the outliers are the ones making the niche product-market choices.

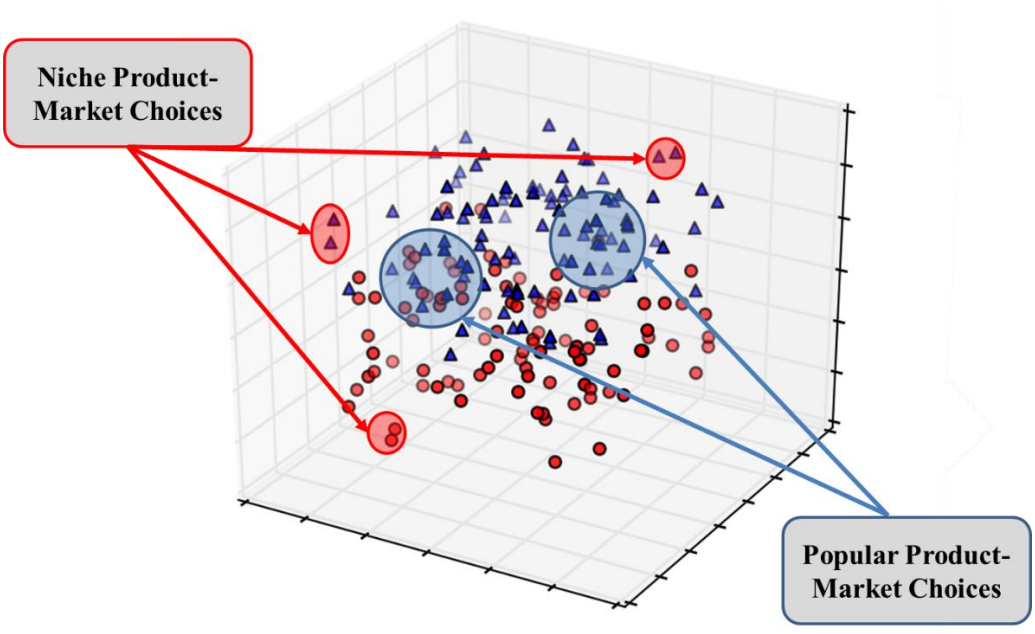


Figure 4.1 Plotting the Firms in the Product-Market Space

For the classification, I use the Mahalanobis Distance. I prefer the Mahalanobis Distance over other distance measure since it is unitless and scale-invariant thereby accounting for the fact that not all firms in the sample will be of the same size. Having classified the firms into niche and popular product-market choices I identify firms which have changed choices over time. Figure 4.2 shows the number of firm-year observations in each group. Of particular interest for the causal inference in this chapter are treatment groups 1 and 2 where the firms have changed choices.

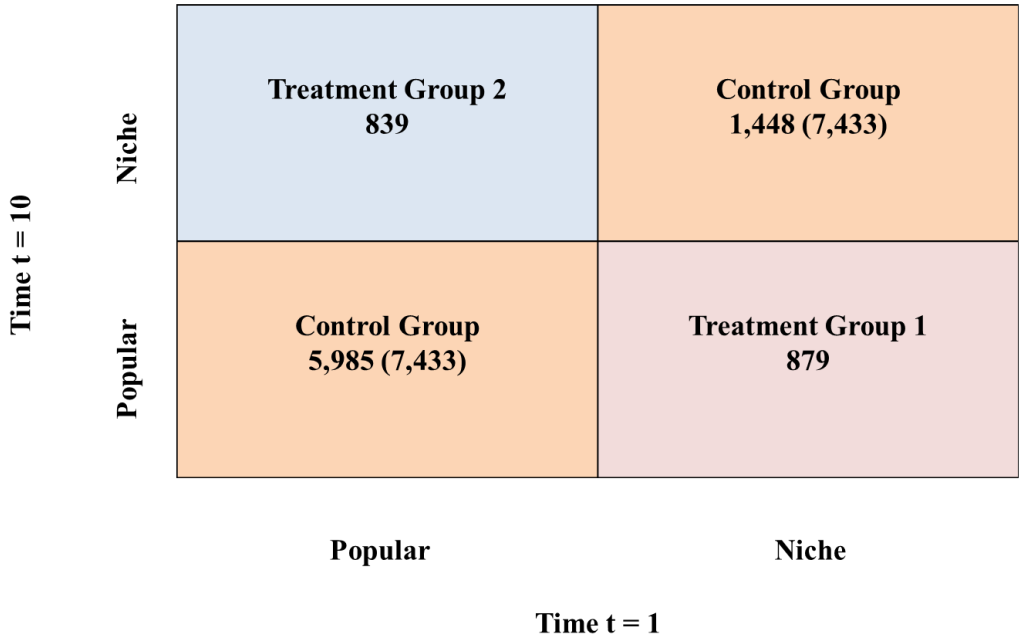


Figure 4.2 Firm Classification based on Product-Market Choices

4.3.2 Matching

One of the challenges in estimating the causal effects for such a diverse sample is the choice of treatment and control groups. Also, these groups need to be comparable to begin with in order to replicate an experimental scenario. In addition, the issue of firms self-selecting themselves into one of the groups also needs to be addressed. This is where researchers have used matching, propensity score matching (PSM) in particular. These matched groups resemble each other before

the treatment firm's decision to switch product-market choices, which creates a statistical equivalence between the groups (Rosenbaum and Rubin 1985).

A majority of the studies using propensity scores to control for imbalances compare only two treatment groups of interest (e.g., treatment and control). However, many researchers have shown that propensity score methods can be extended to cases with three or more conditions of interest (e.g., treatment A, treatment B, and control; (Imbens, 2009; Imai & vanDyk, 2004; Frölich, 2004)). Harder, Stuart & Anthony (2010), Lee, Lessler & Stuart (2011), and Ramchand et al. (2011) in their studies of propensity score estimation of two treatments show that, in terms of bias reduction and mean-squared error, machine learning methods outperform simple logistic regression models with iterative variable section. Extending these findings, machine learning methods may also be advantageous in the multiple treatments setting. The Generalized Boosted Model (GBM) is one such machine learning technique that has been frequently utilized in the two-treatment case (Lee, Lessler & Stuart, 2011; McCaffrey, Ridgeway & Morral, 2004; Ramchand et al., 2011). GBM estimates the propensity score using a flexible estimation method that can adjust for a large number of pretreatment covariates. The estimation involves an iterative process with multiple regression trees. This helps to capture complex and nonlinear relationships between treatment assignment and the pretreatment covariates without over-fitting the data (McCaffrey, Ridgeway & Morral, 2004; Ridgeway, 1999 & 2011; Friedman, 2001 & 2002). GBM not only works with continuous and discrete pretreatment variables but is also invariant to monotonic transformations of them. One of the most useful features of GBM for propensity score estimation is that its iterative procedure can be tuned to find the propensity score model leading to the best balance between treated and control groups. We follow the method outlined in McCaffrey et al. (2013) to estimate the GBM based propensity scores where the algorithm iteratively estimates the weights for each of the groups till

a best balance is achieved. The matched samples are then considered as participants in a survey and on regressing the matched weights with the outcome variable allows us the estimate the effect sizes.

4.4 Data

The data for this analysis comes from all the publicly listed CPG firms in the US. For our analysis, we restrict the sample for the years 2005 – 2014. The financial data comes from Compustat and the text data comes from the firms' 10-K filings with the SEC. We use the Marketing dictionary published by the AMA to filter text to retain only those sentences in the 10-K which contain marketing words from the dictionary. This leaves us with 59,783 terms (which include the terms that have co-occurred with the marketing words) for the 10,331 firm year observations. Table 4.1 provides a full descriptive summary of the variables used in the analysis.

Group	Variable	Unit	Mean	Std Dev	Min	Max
Panel A						
IV & DV	Earnings Per Share	$	228.81	19,414.80	-50,369.00	1,758,000.00
	CapEx	$ Millions	203.45	1,152.56	-11.48	40,595.29
	# Employee	'000	16.43	68.78	0.00	2,200.00
	Market Value	$ Millions	3,667.60	14,835.09	0.00	274,315.44
	Cash Flow	$Millions	-0.31	11.83	-1,273.00	66.43
	Marketing Expense	$ Millions	0.90	11.79	-40.38	825.00
	R & D Expenses	$ Millions	0.05	2.76	0.00	333.33
Panel B						
TDM Size		Terms x Firm Year		59,783 x 10,331		

Table 4.2 Summary Statistics

4.4 Results and Discussion

Following equation (2), we test for the following empirical model:

$$MV_{it} = \boldsymbol{\beta^{(1)}NG\_Change_i} + \beta^{(2)}M\_Dist_{it} + \beta^{(3)}M\_Dist_{it-1} + \beta^{(4)}IND\_MV_{it} + \beta^{(5)}IND\_MV_{it} + \beta^{(6)}MV_{it-1} + \boldsymbol{\beta^{(7)}X_{it}} + \boldsymbol{\beta^{(8)}X_{it-1}} + \varepsilon$$

Where:

MV is the market value of the firm

NG_Change is the categorial variable classifying the firms into one of the three groups

M_Dist is the Mahalanobis distance for the firm in the Product-Market Space

IND_MV is the average market value of all the firms in the industry for the particular year

X includes control variables related to the firm (Size, Growth Opportunities, Industry and Time controls, B2B vs B2C classification for firms)

Table 4.3 shows the results of the analysis[4]. We can see for the results that as a result of the matching exercise, we lose a few firm-year observations for each of the groups, but the resulting matched sample is the one with the best balance. The top panel suggests that the average effect for a firm moving towards niche product – market choices is significant and positive when controlling for the firms that have not changed strategies as well as firms that have switched to popular product-market choices. The result continues to hold the same direction (although the significance levels drop from 5% to the 10% level) when additional variables are used in the econometric estimation to control for the past performance.

[4] Other estimation results are available in the working paper based on this chapter.

No Lagged Variables included

Sample sizes and effective sample sizes:

treatment	n	ESS.es.mean	ESS:ks.mean
1	5878	5878	5878
2	783	783	783
3	783	783	783

Treatment Effects

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**5.43692**	**0.03127**	**173.861**	**<2e-16**	*******
Treatment Group 1	-0.1322	0.09334	-1.416	0.1567	
Treatment Group 2	**0.20163**	**0.08924**	**2.259**	**0.0239**	*****

With Lagged Variables included

Sample sizes and effective sample sizes:

treatment	n	ESS.es.mean	ESS:ks.mean
1	3873	3873	3873
2	475	475	475
3	462	462	462

Treatment Effects

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**5.65178**	**0.03425**	**165.019**	**<2e-16**	*******
Treatment Group 1	-0.06352	0.10734	-0.592	0.554	
Treatment Group 2	***0.18024***	***0.10429***	***1.728***	***0.084***	**.**

Table 4.3 Regression Results

4.5 Conclusion

To summarize, we started by asking the question that Can the seemingly open-ended product-market choice space for tech sector firms be adequately represented in a condensed, finite dimensional space? In particular, we wanted to understand how this can be achieved. We argued the presence of a product-market choices space which is latent in the senior management's cognition. This allowed us to reduce the unstructured textual data from the 10-K filings into manageable metric variables (product-market choices) using the topic modeling technique we proposed in chapter 2. We next plotted the firms in the multi-dimensional product-market space and then identify the outliers, labelling them as firms with niche product-market choices. We then study the causal relationship by identifying those firms who have switched from niche product-market choices to popular choices and vice versa (*treatment groups*), with those firms who haven't changed the choices (*control group*). Using a GBM based PSM for multiple treatments, we estimate the effect size of the strategic change to find a significant and positive effect for those firms which have switched from popular product-market choices to niche product-market choices in the timeframe we have considered.

4.6 Limitations and Future Research

This research has certain limitations which future research can address. This chapter only considers the case when firms have switched choices once I the time period. However, future research could look at choices at intermediate years (for example, every 2 years) and better explain the causal relationship. Also, the chapter considers that the product-market choices arrived at using the text mining algorithm are independent of each other. Future research could look at clustering such choices to provide more actionable insights for managers.

Chapter 5 – Conclusion

In this dissertation I used a marketing lens to measure the firm related communications from unstructured data to better understand its implications for consumers, firms and regulators. I defined firm related communications broadly to include all types of discussion captured in unstructured format- consumer opinions as captured in consumer reviews and blogs, intra firm discussions and, public communications by firms and regulatory institutions. I combined econometric methods and natural language technologies to examine the influences on and of firm related communications. The first of the three essays proposed and validated an extension to a popular latent topic modeling algorithm which enabled better interpretation of consumer insight from these communications. The second essay used this proposed extension to derive a measure of marketing effectiveness from communications and used it to explain brand performance. The third essay used this proposed extension to classify firms as following a niche or popular product-market choices based on strategic information in their statutory filings. I found a statistically significant performance differential attributable to product-market choices.

As shown in figure 5.1, dissertation started with considering all forms of firm and brand related communications. I applied text analytic methods on this unstructured textual data to derive metric variables with marketing insights. These metric variables then aided explaining firm and brand performance with actionable insights for managers. The insights are expected to result in updated communication strategies which will create updated textual data and the cycle is expected to continue.

Figure 5.1 Applying a Marketing Lens on Firm Related Communications

References

Abrahamson, E., & Hambrick, D. C. (1997). Attentional homogeneity in industries: The effect of discretion. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *18*(S1), 513-532.

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In *International encyclopedia of statistical science* (pp. 206-208). Springer, Berlin, Heidelberg.

Alpaydin, E. (2010). *Introduction to machine learning*. MIT press.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science, 57*(8), 1485–1509.

Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of the literature. *Research Policy, 16*(2–4), 131–141.

Baum, J. R., Locke, E. A., & Kirkpatrick, S. A. (1998). A longitudinal study of the relation of vision and vision communication to venture growth in entrepreneurial Firms. *Journal of Applied Psychology*, *83*(1), 43.

Bharadwaj, A. S., Bharadwaj, S. G., & Konsynski, B. R. (1999). Information technology effects on firm performance as measured by Tobin's q. *Management Science, 45*(7), 1008–1024.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of*

the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). AcM.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Bracker, K., & Ramaya, K. (2011). Examining the impact of research and development expenditures on Tobin's Q. *Academy of Strategic Management Journal*, *10*, 63.

Bronzini, R., & Piselli, P. (2016). The impact of R&D subsidies on firm innovation. *Research Policy*, *45*(2), 442-457.

Bruce, N. I., Foutz, N. Z., & Kolsarici, C. (2012). Dynamic effectiveness of advertising and word of mouth in sequential distribution of new products. *Journal of Marketing Research*, *49*(4), 469-486.

Brynjolfsson, E., Hitt, L. M., and Yang, S. K. 2002. Intangible Assets: Computers and Organizational Capital, *Brookings Papers on Economic Activity* (1), pp. 137-198.

Brynjolfsson, E., & Hitt, L. (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management science*, *42*(4), 541-558.

Cassiman, B., & Veugelers, R. (2006). In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. *Management Science*, *52*(1), 68–82.

Cho, T. S., & Hambrick, D. C. (2006). Attention as the mediator between top management team characteristics and strategic change: The case of airline deregulation. *Organization Science*, *17*(4), 453-469.

Clapham, S. E., & Schwenk, C. R. (1991). Self-serving attributions, managerial cognition, and company performance. *Strategic Management Journal*, *12*(3), 219-229.

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*(1), 37-56.

Collins James, C., & Porras, J. I. (1991). Organizational Vision and Visionary Organization. *California Management Review*, *34*(1), 30–52.

Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115-122). ACM.

Daly, J. P., Pouder, R. W., & Kabanoff, B. (2004). The effects of initial differences in firms' espoused values on their postmerger performance. *The Journal of Applied Behavioral Science*, *40*(3), 323-343.

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

Dechow, Patricia M. (1994), Accounting earnings and cash flows as measures of firm performance: The role of accounting accruals, *Journal of Accounting and Economics*, 18 (1), 3–42.

Dechow, Patricia M., Sagar P. Kothari, and Ross L. Watts (1998), The relation between earnings and cash flows, *Journal of Accounting and Economics*, 25 (2), 133–168.

de Vries, L., Gensler, S., & Leeflang, P. S. (2017). Effects of traditional advertising and social messages on brand-building metrics and customer acquisition. *Journal of Marketing*, *81*(5), 1-15.

Doyle, P. (1975). Brand positioning using multidimensional scaling. *European Journal of Marketing*, *9*(1), 20-34.

Dutta, Shantanu, Om Narasimhan, and Surendra Rajiv (1999), Success in high-technology markets: Is marketing capability critical? *Marketing Science*, 18 (4), 547–568.

Eggers, J. P., & Kaplan, S. (2009). Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change. *Organization Science*, *20*(2), 461-477.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, *17*(1), 43-57.

Feng, H., Morgan, N. A., & Rego, L. L. (2015). Marketing department power and firm performance. *Journal of Marketing*, *79*(5), 1-20.

Filion, L. J. (1991). Vision and Relations: Elements for an Entrepreneurial Metamodel. *International Small Business Journal*, *9*(2), 26–40.

Fiol, C. M. (1995). Corporate communications: Comparing executives' private and public statements. *Academy of Management Journal*, *38*(2), 522-536.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378.

Frölich, M. (2004). Programme evaluation with multiple treatments. *Journal of Economic Surveys*, *18*(2), 181-224.

Ganesan, K., Zhai, C., & Han, J. (2010, August). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*(pp. 340-348). Association for Computational Linguistics.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Germann, F., Ebbes, P., & Grewal, R. (2015). The chief marketing officer matters!. *Journal of Marketing*, *79*(3), 1-22.

Gerrish, S., & Blei, D. M. (2010, June). A Language-based Approach to Measuring Scholarly Impact. In *ICML* (Vol. 10, pp. 375-382).

Gilbert, E., & Karahalios, K. (2010, May). Widespread Worry and the Stock Market. In *ICWSM* (pp. 59-65).

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012.

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

Grant RM. (2008). Contemporary Strategy Analysis. Blackwell Pub.

Griliches, Z. (1990). *Patent Statistics as Economic Indicators: A Survey* (Working Paper No. 3301). National Bureau of Economic Research.

Hanssens, D. M., & Pauwels, K. H. (2016). Demonstrating the value of marketing. *Journal of Marketing*, *80*(6), 173-190.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3), 234.

Helmers, C., & Rogers, M. (2011). Does patenting help high-tech start-ups?. *Research Policy*, *40*(7), 1016-1027.

Henderson, B., Kobelsky, K. W., Richardson, V. J., and Smith, R. 2010. The Relevance of Information Technology Expenditures, *Journal of Information Systems* (24:2), pp. 39-77.

Hoberg, G., & Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, *23*(10), 3773–3811.

Hoberg, G., Phillips, G. M., & Prabhala, N. R. (2013). Product market threats, payouts, and financial flexibility. *The Journal of Finance*, *69*(1), 8.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854-866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706-710.

Jacobs BJD, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. *Marketing Science*. 35(3):389–404

Khanna, T., & Palepu, K. (2000). Is group affiliation profitable in emerging markets? An analysis of diversified Indian business groups. *Journal of Finance*, 867–891.

Kim, Myungsun, and William Kross (2005), The ability of earnings to predict future operating cash flows has been increasing-not decreasing, *Journal of Accounting Research*, 43 (5), 753–80.

King, G., & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, *57*(3), 617-642.

Kirkpatrick, S. A., Wofford, J. C., & Baum, J. R. (2002). Measuring motive imagery contained in the vision statement. *The Leadership Quarterly*, *13*(2), 139–150.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009, May). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 272-280). Association for Computational Linguistics.

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*(1), 7-25.

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*(2), 194-212.

Lang, L., Ofek, E., & Stulz, R. (1996). Leverage, investment, and firm growth. *Journal of Financial Economics*, *40*(1), 3–29.

Larwood, L., Falbe, C. M., Kriger, M. P., & Miesing, P. (1995). Structure and Meaning of Organizational Vision. *The Academy of Management Journal*, *38*(3), 740–769.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000, November). Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management* (pp. 389-396). ACM.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), 337-346.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, *6*(3), e18174.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

McAlister, Leigh, Raji Srinivasan, and MinChung Kim (2007), Advertising, research and development, and systematic risk of the firm, *Journal of Marketing*, 71 (1), 35–48.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, *32*(19), 3388-3414.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, *27*(1), 57-86.

Menon, N. M., Lee, B., and Eldenburg, L. 2000. Productivity of Information Systems in the Healthcare Industry, *Information Systems Research* (11:1), pp. 83-92.

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 165-171). IEEE.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*.

Mimno, D., & McCallum, A. (2007, August). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 500-509). ACM.

Mintzberg, H. (1978). Patterns in strategy formation. *Management Science*, *24*(9), 934–948.

Mithas, S., Tafti, A., Bardhan, I., and Goh, J. M. 2012. Information Technology and Firm Profitability: Mechanisms and Empirical Evidence, *MIS Quarterly* (36:1), pp. 205-224.

Mitsuhashi, H., & Greve, H. R. (2009). A matching theory of alliance formation and organizational success: Complementarity and compatibility. *Academy of Management Journal*, *52*(5), 975-995.

Nath, P., & Mahajan, V. (2008). Chief marketing officers: A study of their presence in firms' top management teams. *Journal of Marketing*, *72*(1), 65-81.

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

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm*, *11*(122-129), 1-2.

Osborne, J. D., Stubbart, C. I., & Ramaprasad, A. (2001). Strategic groups and competitive enactment: A study of dynamic relationships between mental models and performance. *Strategic Management Journal*, *22*(5), 435-454.

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *Icwsm*, *20*, 265-272.

Pauwels, K., Demirci, C., Yildirim, G., & Srinivasan, S. (2016). The impact of brand familiarity on online and offline media synergy. *International Journal of Research in Marketing*, *33*(4), 739-753.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Petersen, J. A., Kumar, V., Polo, Y., & Sese, F. J. (2017). Unlocking the power of marketing: Understanding the links between customer mindset metrics, behavior, and profitability. *Journal of the Academy of Marketing Science*, 1-24.

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ramage, D., Manning, C. D., & Dumais, S. (2011, August). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457-465). ACM.

Ramchand, R., Griffin, B. A., Suttorp, M., Harris, K. M., & Morral, A. (2011). Using a cross-study design to assess the efficacy of motivational enhancement therapy–cognitive behavioral therapy 5 (MET/CBT5) in treating adolescents with cannabis-related disorders. *Journal of studies on alcohol and drugs*, *72*(3), 380-389.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, 172-181.

Ridgeway, G. (2011). Generalized boosted regression models: A guide to the gbm package. *2007 08-03)[2014 09-30]. http,//ftp.ctex.org/mirrors/cran/web/packages/gbm*.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Schilling, M. A., & Phelps, C. C. (2007). Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation. *Management Science, 53*(7), 1113–1126.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, *12*(4), 561-587.

Shellman, S. M. (2008). Coding disaggregated intrastate conflict: machine processing the behavior of substate actors over time and space. *Political Analysis*, *16*(4), 464-477.

Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. *MIS Quarterly*, *40*(4).

Silber, J. H., Rosenbaum, P. R., & Ross, R. N. (1995). Comparing the Contributions of Groups of Predictors: Which Outcomes Vary With Hospital Rather Than Patient Characteristics. *Journal of the American Statistical Association*, *90*(429), 7–18.

Taddy, M. A. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, *108*(503), 755–770.

Taddy, M. A. (2011). On Estimation and Selection for Topic Models. *arXiv:1109.4518 [stat]*.

Tanriverdi, H., & Venkatraman, N. (2005). Knowledge relatedness and the performance of multibusiness firms. *Strategic Management Journal*, *26*(2), 97.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139-1168.

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*(4), 463–479.

Wang, L., & Zajac, E. J. (2007). Alliance or acquisition? A dyadic perspective on interfirm resource combinations. *Strategic management journal*, *28*(13), 1291-1317.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 178–185). ACM.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Winter, D. G., & McClelland, D. C. (1978). Thematic analysis: An empirically derived measure of the effects of liberal arts education. *Journal of Educational Psychology*, *70*(1), 8.

Worm, S., Bharadwaj, S., Shen, J., & Srivastava, R.K., (2016). Marketing Does Matter in The Boardroom: Firm Performance Outcomes of Top-Management-Team Marketing Mindset. *Working Paper*

Xiong, G., & Bharadwaj, S. (2013). Asymmetric roles of advertising and marketing capability in financial returns to news: Turning bad into good and good into great. *Journal of Marketing Research*, 50(6), 706-724.

Yadav, M. S., Prabhu, J. C., & Chandy, R. K. (2007). Managing the future: CEO attention and innovation outcomes. *Journal of Marketing*, *71*(4), 84-101.

Yermack, D. (1996). Higher market valuation of companies with a small board of directors. *Journal of Financial Economics*, *40*(2), 185–211.

APPENDIX A - Extending the LDA Approach for Improved Qualitative Analyses on Text Data

Figure A.1.a Word Cloud and Co-occurrence Graph for Latent Topic 1 – Study 2

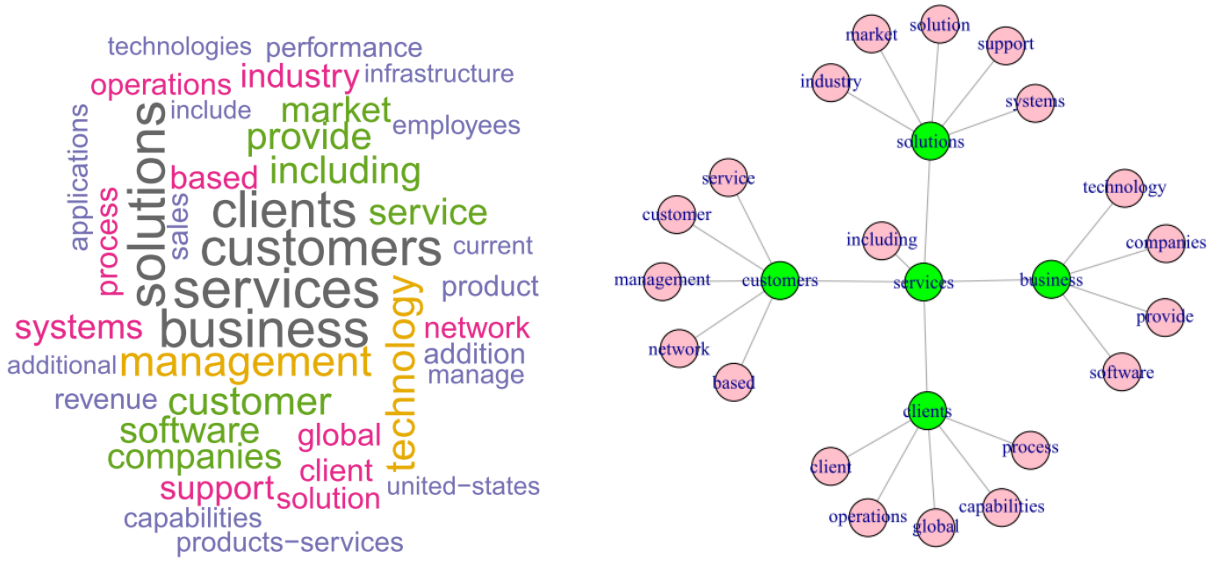


Figure A.1.b Word Cloud and Co-occurrence Graph for Latent Topic 2 – Study 2

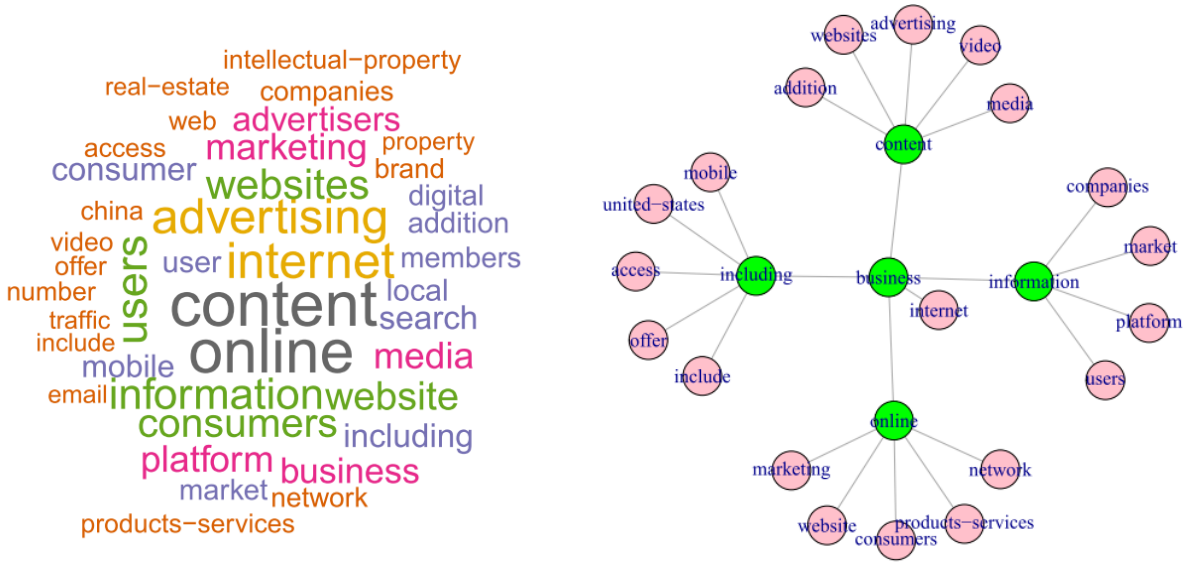


Figure A.1.c Word Cloud and Co-occurrence Graph for Latent Topic 3 – Study 2

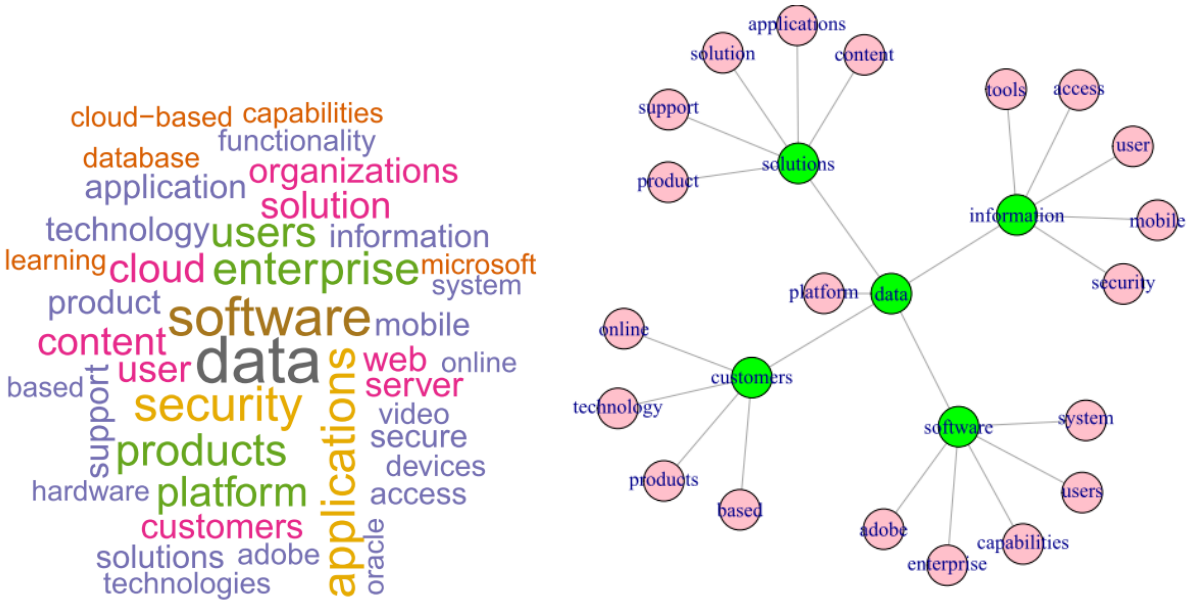


Figure A.1.d Word Cloud and Co-occurrence Graph for Latent Topic 4 – Study 2

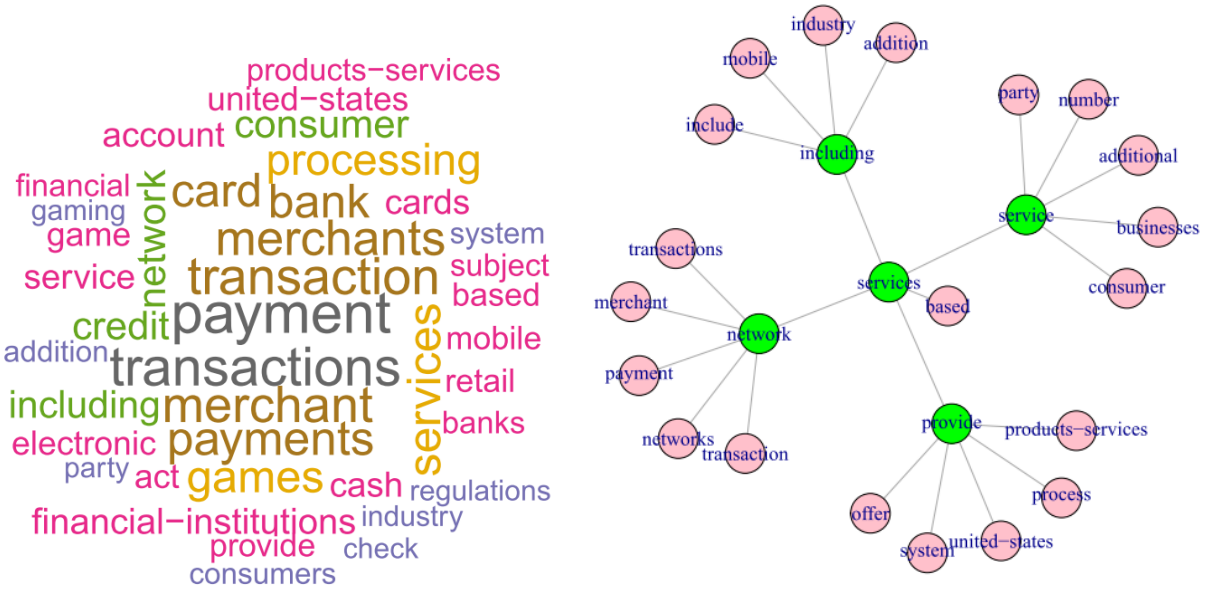


Figure A.1.e Word Cloud and Co-occurrence Graph for Latent Topic 5 – Study 2

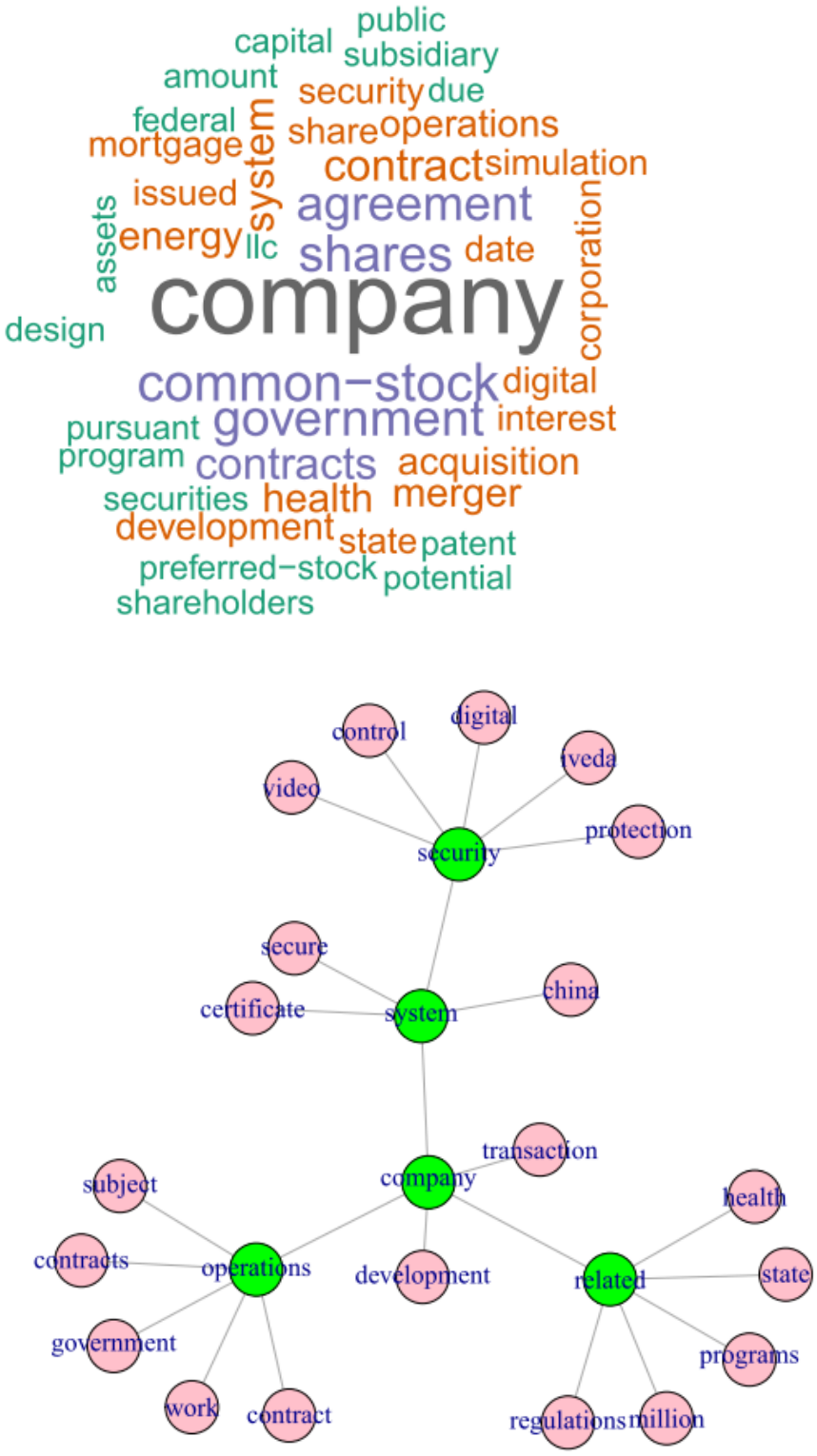


Table A.2. Full Table of Regression Results – Study 2

	Dependent variable: log (EPS)	
	Lag Only	**With Concurrent**
Controls		
EPS $_{t-1}$	**-0.071 (0.028)**	**-0.070 (0.028)**
	p = 0.013**	**p = 0.015****
Cash Flow $_{t-1}$ [a]	0.010 (0.014)	0.011 (0.014)
	p = 0.467	p = 0.461
Marketing Expenses $_{t}$ [a]	-0.013 (0.025)	-0.011 (0.025)
	p = 0.599	p = 0.670
Marketing Expenses $_{t-1}$ [a]	-0.008 (0.021)	-0.008 (0.022)
	p = 0.714	p = 0.720
R&D Expenses $_{t}$ [a]	-0.0004 (0.029)	-0.002 (0.029)
	p = 0.990	p = 0.956
R&D Expenses $_{t-1}$ [a]	0.014 (0.022)	0.013 (0.022)
	p = 0.538	p = 0.546
Growth Opportunity $_{t}$	-0.008 (0.037)	-0.009 (0.037)
	p = 0.829	p = 0.816
RoA $_{t}$	**0.106 (0.019)**	**0.107 (0.020)**
	p = 0.00000***	**p = 0.00000*****
# Employees $_{t}$	0.168 (0.164)	0.168 (0.165)
	p = 0.304	p = 0.308
Liabilities $_{t}$ [a]	-0.026 (0.058)	-0.028 (0.059)
	p = 0.653	p = 0.638
Topic Factors		
B2B Solutions Focus $_{t}$		0.005 (0.125)
		p = 0.971
Online Content Focus $_{t}$		-0.023 (0.127)
		p = 0.859
Data & App Management Capbilities $_{t}$		-0.018 (0.129)
		p = 0.891
Transaction Enabling Capabilities $_{t}$		-0.060 (0.134)
		p = 0.653
B2B Solutions Focus $_{t-1}$	**0.267 (0.103)**	**0.260 (0.105)**
	p = 0.010***	**p = 0.014****
Online Content Focus $_{t-1}$	**0.278 (0.100)**	**0.280 (0.102)**
	p = 0.006***	**p = 0.007*****
Data & App Management Capbilities $_{t-1}$	***0.197 (0.117)***	0.194 (0.121)
	p = 0.093*	p = 0.110
Transaction Enabling Capabilities $_{t-1}$	**0.189 (0.070)**	**0.191 (0.072)**
	p = 0.008***	**p = 0.009*****
Intercept	-0.036 (0.309)	-0.012 (0.355)
	p = 0.908	p = 0.973
Observations	1,606	1,606
R^2	0.87	0.87
Adjusted R^2	0.804	0.804
Residual Std. Error	0.442 (df = 1066)	0.443 (df = 1062)
F Statistic	13.250*** (df = 539; 1066)	13.112*** (df = 543; 1062)

Note:All variables are log transformed; [a] *Scaled by Total Assets* *p<0.1; **p<0.05; ***p<0.01

APPENDIX B - Brand Marketing Effectiveness, Brand Positioning Gaps and Brand Outcomes: An Empirical Investigation

Property	Measure
Variability	Measure of Textual Lexical Diversity (MTLD)
Volume	Total number of words in the text
Evenness	Standard deviation of tokens per type
Rarity	Mean BNC rank
Dispersion	Mean distance between tokens of type
Disparity	Mean number of words per sense or Latent Semantic Analysis

SUPPLEMENTARY ANNEXURE - Extending the LDA Approach for Improved Qualitative Analyses on Text Data – Survey Questionnaire

Supplementary Annexure / Web Appendix

Survey Questionnaires

(***Note**: All surveys were administered online on Qualtrics*)

Survey 1. LDA Topic Interpretation for Samsung Galaxy Reviews

Q1 Recall the Nokia Lumia example from class? We topic-mined consumer reviews of Nokia Lumia 925 and identified two latent topics underlying the text. We labeled the topics and interpreted what they likely meant.We used the word-clouds and co-occurrence graphs as visual aids for topic identification. Recall that font-size in a word-cloud is proportional to word frequency in the text. In a co-occurrence graph, words that likely occur together within a document are connected in a network of the words. Aside from that, we also used the top few documents loading onto each topic to help in topic interpretation.The following survey task asks you to repeat the Nokia type exercise on a new mobile brand - Samsung Galaxy.Below, you will be presented with the topic-mining results for Samsung Galaxy reviews (collected around the same time as Nokia's). Two topics were found to be optimal. You'll be shown word-clouds and co-occurrence graphs for each of the topics. Further, you'll also be shown the top 5 of the documents that loaded on each topic (as a separate file on linkstreet). Your task is (i) to interpret these topics using the output, (ii) to meaningfully label these two topics with a short, informative name and (iii) provide a brief description of your interpretation. Remember, there is no right or wrong answer, just your carefully considered responses.Sudhir

Q12 Your full name

Q13 Your registration number

Q2

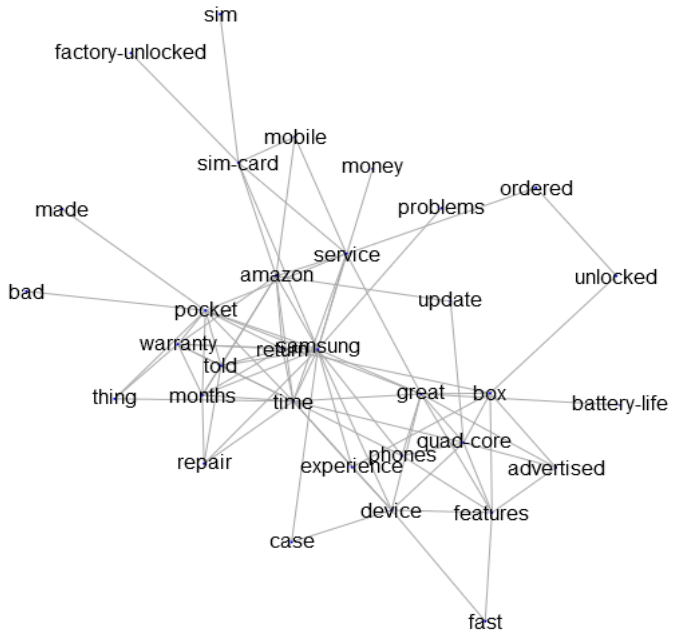
Q3

Q4 Based on the word-cloud and the co-occurrence graph, informatively label the topic (not more than 4-6 words)

Q8 Please provide a brief justification for your choice of the topic name.

Q5

Q6



Q9 Based on the wordcloud and the co-occurence graph, please provide a name for the topic (not more than 4-6 words)

Q10 Please provide a brief justification for your choice of the topic name.

Survey 2. LDA Token-topic validation for Nokia Lumia and Samsung Galaxy Reviews

This survey aims to assess machine classification of terms when pitted against a superior human standard of classification. Please remember that there are no right or wrong answers in what follows. I might use an anonymized and aggregate form of this data for academic research (i.e., if and only if the analysis results are interesting from a research standpoint). Please be assured of strict data confidentiality. **NO** part of your personal data or identity will at any stage be visible to anyone but me. However, if you still have concerns regarding your participation in the survey, please let me know.

While text-analyzing consumer reviews on Amazon on the Nokia Lumia 900 smartphone, two main themes or topics of interest emerged from the data. One topic related to operating system (or OS) related concerns/feedback. I labeled it "OS Specific". Note that the Lumia had a relatively unknown Windows mobile OS when it was introduced. The second topic seemed to concern a discussion of features of the phone such as the camera, touch-screen, battery etc. In other words, it did not seem to directly concern the OS. Hence, I labelled it "OS Neutral"

Next you will be presented with a random sampling of 25 tokens. Please classify them into either "OS Specific" or "OS Neutral".

Please classify the tokens

	OS Specific (1)	OS Neutral (2)
nokia (1)	❍	❍
windows (2)	❍	❍
apps (3)	❍	❍
great (4)	❍	❍
good (5)	❍	❍
phones (6)	❍	❍
battery (7)	❍	❍
don (8)	❍	❍
android (9)	❍	❍
time (10)	❍	❍
music (11)	❍	❍
app (12)	❍	❍
love (13)	❍	❍
screen (14)	❍	❍

work (15)	❍	❍
back (16)	❍	❍
iphone (17)	❍	❍
camera (18)	❍	❍
easy (19)	❍	❍
mobile (20)	❍	❍
user (21)	❍	❍
find (22)	❍	❍
people (23)	❍	❍
price (24)	❍	❍
contacts (25)	❍	❍
unlocked (26)	❍	❍
works (27)	❍	❍
cell (28)	❍	❍
bought (29)	❍	❍
seller (30)	❍	❍
day (31)	❍	❍
device (32)	❍	❍
product (33)	❍	❍
didn (34)	❍	❍
excellent (35)	❍	❍
buy (36)	❍	❍
happy (37)	❍	❍

features (38)	❍	❍
software (39)	❍	❍
bluetooth (40)	❍	❍
email (41)	❍	❍
sim (42)	❍	❍
view (43)	❍	❍
awesome (44)	❍	❍
text (45)	❍	❍
box (46)	❍	❍
sms (47)	❍	❍
update (48)	❍	❍
list (49)	❍	❍
amazon (50)	❍	❍

In a similar study for text-analyzing consumer reviews on Amazon on the Samsung Galaxy S3 smartphone, two main themes or topics of interest emerged from the data. One, topic related to warranty and service related issues/feedback. I labeled it "Warranty & Service". The second topic related to general features discussion of the phone like camera, apps, touch-screen, battery etc. I labeled it labelled "Features discussion (both OS and non-OS)". Next you will be presented with a random sampling of 25 tokens. Please classify them into the topic you think it most likely belongs in.

Please classify the tokens

	Warranty & Service (1)	Features Discussion (both OS & non-OS) (2)
version (1)	❍	❍
unlocked (2)	❍	❍
great (3)	❍	❍
mobile (4)	❍	❍
sim (5)	❍	❍
service (6)	❍	❍
fast (7)	❍	❍
don (8)	❍	❍
doesn (9)	❍	❍
box (10)	❍	❍
works (11)	❍	❍
features (12)	❍	❍
phones (13)	❍	❍
card (14)	❍	❍
factory (15)	❍	❍
told (16)	❍	❍
life (17)	❍	❍
warranty (18)	❍	❍
device (19)	❍	❍
made (20)	❍	❍
return (21)	❍	❍

core (22)	❍	❍
advertised (23)	❍	❍
iii (24)	❍	❍
days (25)	❍	❍
samsung (26)	❍	❍
screen (27)	❍	❍
galaxy (28)	❍	❍
battery (29)	❍	❍
amazon (30)	❍	❍
bought (31)	❍	❍
good (32)	❍	❍
buy (33)	❍	❍
seller (34)	❍	❍
love (35)	❍	❍
product (36)	❍	❍
work (37)	❍	❍
iphone (38)	❍	❍
day (39)	❍	❍
camera (40)	❍	❍
months (41)	❍	❍
people (42)	❍	❍
recommend (43)	❍	❍
make (44)	❍	❍

android (45)	❍	❍
purchase (46)	❍	❍
apps (47)	❍	❍
problem (48)	❍	❍
voice (49)	❍	❍
amazing (50)	❍	❍

Survey 3. Document Classification for Nokia Lumia and Samsung Galaxy Reviews
Dear Class, This survey aims to assess machine classification of terms and 'documents' when pitted against a superior human standard of classification. There are no right or wrong answers in what follows, so just diligently solve the questions posed to you. This homework is graded on timeliness and completeness. There is a possibility that I might use an anonymized and aggregate form of this data for academic research (i.e., if and only if the analysis results are interesting from a research standpoint). Please be assured of strict data confidentiality. **NO** part of your personal data or identity will at any stage be visible to anyone but me. However, if you still have concerns regarding your participation in the survey, please let me know.

While text-analyzing consumer reviews on Amazon on the Nokia Lumia 900 smartphone, two main themes or topics of interest emerged from the data. One topic related to operating system (or OS) related concerns/feedback. I labeled it "OS Specific". Note that the Lumia had a relatively unknown Windows mobile OS when it was introduced. The second topic seemed to concern a discussion of features of the phone such as the camera, touch-screen, battery etc. In other words, it did not seem to directly concern the OS. Hence, I labelled it "OS Neutral" Next you will be presented with a random sampling of ten review excerpts. Please read them. Classify them based on their orientation into either "OS Specific" or "OS neutral.

Please classify the reviews based on which topic you think it belongs.

	OS Specific (1)	OS Neutral (2)
great peice easy to handle and i like it very mcuh this is my third gift to my friend pefect (162)	❍	❍
if only all phones were nokia then the world would be a happier place no other phone for me great (163)	❍	❍
this product have an att app in my country dont have att services how can change this or delete this (164)	❍	❍
it is amazing just wish had more apps but otherwise its prefect it is alot faster then my old apple iphone (165)	❍	❍
it connected me with my family i am using it without any problem the best is the camera which is mega pixels (166)	❍	❍
the phone is everything that i expected it would be and it s easy to use for even the inexperienced cell phone user (167)	❍	❍
i love this phone it works great it s fast and the sleek look is awesome i love the color windows phones are awesome (168)	❍	❍

it s the best phone i ve ever bought so far i recommend to anyone who is looking for a good phone for a affordable price (169)	❍	❍
costumers realy have to like windows phone os and this phone its a good way of have everithing synchronyze with your computer (170)	❍	❍
beautiful color very nice keys wish it had removable mem card but is very fluent hasnt hung up yet have the tuxedo and screen protectors on it (171)	❍	❍
great phone used an iphone before and i am much happier now with the lumia i did not had a tousand apps in my iphone so the transition was nice and easy (172)	❍	❍
this is totally a unlocked phone you can use it anywhere this is a great phone for anyone who is interested in windows phone it is much better than iphone (173)	❍	❍
the phone works well and is pretty light which is greatly appreciated having previously owned an android phone i was disappointed by the lack of some apps i had gotten used too but everything else is	❍	❍

pretty great (174) i bought this phone because i heard so many great things about it however it does not allow many apps at all quite disappointing since i love instagram this really broke my heart its a great phone but not if your an app person (175) great phone i went with this phone to replace my htc trophy as my windows phone just to let you all know i am using this phone over an iphone s and a nexus this is a rock solid phone and love it windows phone is still a great os no need for wp (176) this was a great smartphone for its time got this as a replacement phone for my mother it is disappoint it doesn t support all of the new apps coming out that only support windows phone lumia is running the app selection is really what cripples this phone (177) because it has the same camera as iphone has i bought but when i take the pictures or videos it can not beat iphone at all especially at night i am a little bit disappointed i used to use a htc and my data worked well by this phone i always	❍ ❍ ❍ ❍	❍ ❍ ❍ ❍

lost my data connection i don t know why it happens (178)		
i really like the phones and i think i will be getting my own soon though i don t know if my parents will find them easy to for they were a gift to my parents the only thing i was ever happy about is the confusion that occurred when i placed this items but i hope it doesn t happen more often with some of your customers thanks for the discount though from the shipping fee (179)	❍	❍
good looking phone sturdy gorgeous screen and unlimited music nokia music free updated gps maps from most of the world the apps i need and a clean interface i don t like spending time customizing the appearance of a phone with this you have to be productive it does the job i can say it is really a steal for the recent price last week it got just better with the upgraded to (180)	❍	❍
excellent brand i have to adapt myself to this kind of technology and this is a good way to start (181)	❍	❍
got this phone last week and really like it although they	❍	❍

unboxed itnow i m waiting for wp version (182)		
i like all nokia products and this is awasome besides my son bought for a gift and it works perfectly (183)	❍	❍
it s not unlocked i am very dissatisfied with this product the main reason to buy it is that was unlocked (184)	❍	❍
unlocked i can use it for text and call and i do have data plan but it sd not working was i doing something wrong (185)	❍	❍
beautiful device works fine and boy it has great audio quality and video its dial is easy to read i am happy with it (186)	❍	❍
i totally loved the phone but unfortunately it was locked but i was looking for an unlocked phone i am very disappointed by this situation (187)	❍	❍
thanks for all the nokia lumia i arrived on time and in my house i like it is a excellent condition and as a customer i m happy thank you very much (188)	❍	❍
as you promise the item arrive on time thanks i have to say	❍	❍

that i was able to open the box a few days later because a wasn t at home the day o arrival (189)		
very nice excellent cell phone access is highly recommended fast internet has many applications to download and it works very well the truth is excellent buy (190)	❍	❍
great phone full of features and visually appealing menus although gsm unlocked unable to get data across straight talk apn and wifi suspect data would work across at t network (191)	❍	❍
product arrived on time so far i have had no problems with the product seems robust like i expected from nokia i have never owned another brand and this hasn t changed my preference yet (192)	❍	❍
overall a good phone not a big fan of the app game collection though have to press the back button a hundred times to close out of pages i ve opened or they keep running its got zune instead of itunes (193)	❍	❍
the only thing i received in my phone is the phone and a house charger i don t know	❍	❍

how to open the phone or how to operate the phone can you send me the rest of the stuff to go along with the phone please (194)		
very nice and bright display running very smoothly unfortunately still i have not connect it to internet due to the fact it requires mini sim card still i am waiting for it from my european mobile operator (195)	❍	❍
a very good phone connection fast an excellent size especially for viewing websites allows sharing of a simple fast and fluid the downside of this device is that suddenly goes off to turn it back on you have to perform a soft reset (196)	❍	❍
since i received the phone i have not been able to use because it asks me for a code to unlock and in venezuela i have sought someone that knows how to unlock it and nobody knows now i do not know what to do with the phone i feel very cheated (197)	❍	❍
i was very disappointed when i purchased the phone it said new but it wasn t the phone had a couple scratches and it	❍	❍

was repackaged it did not come in a nokia lumia box and the only thing i got was a charger the phone itself works fine but the seller failed (198) i chose this phone because i like the features the problem is when i send multiple text it come back unsend i had to resend it individually also some incoming text i am unable to open because it shows a link to download media contact but when i click on the link it doesn t download help (199)	❍	❍

While text-analyzing consumer reviews on Amazon on the Samsung Galaxy S3 smartphone, two main themes or topics of interest emerged from the data. One, topic related to warranty and service related issues/feedback. I labeled it "Warranty & Service". The second topic related to general features discussion of the phone like camera, apps, touch-screen, battery etc. I labeled it labelled "Features discussion (both OS and non-OS)". Next you will be presented with a random sampling of ten review excerpts. Please read them. For each review, please select with a tick mark which topic you think it most likely belongs in.

Please classify the reviews into the topic you think it belongs

	Warranty & Service (1)	Features discussion (both OS and non-OS) (2)
it s very wellit s new phone i like it very muchthanks give me a nice phone (169)	❍	❍
i liked it as it is i wish it had double sim slots though very fast and slim excellent product (170)	❍	❍
although it appears in the description it is unlocked phone all the software is tied to at t disappointed (171)	❍	❍
this is not an international version only small sim card works doesn t work with all companies i am so frustrated (172)	❍	❍
excellent phone the only thing that is not included es the geotaggig geature in the camera you should point this out (173)	❍	❍
would not accept sim chip as it would not lock into placebroken sim cover overall shoddy workmanship and i would not recommend (174)	❍	❍
excelente producto rapido en envio por liberty express a venezuela y barata y en atencion excelente lo recomiendo excelente calidad	❍	❍

(175) i bought the phone a few days ago it was not an international version and it didn t work in my country it said it was for europe only	❍	❍
(176) was exactly what it said it was came sealed brand new all parts were included in the box even the headphones im happy with the purchase	❍	❍
(177) a great phone with many features but sadly this version doesnt support g t mobile signal ahve g sometimes planning to change to at t though	❍	❍
(178) probably the best phone i ve had in my life despite the fact that sold already the s it is still very much alive very good display storage and performance	❍	❍
(179) this cell is not released it is the opposite of what he says would not recommend este celular no es liberado es todo lo contrario que dice no lo recomiendo	❍	❍
(180) it seems an excellent phone it has very good applications it is very light and very fast connected with the internet the	❍	❍

supplier fulfilled with the delivery time (181)		
these people sell cell clones koreans are not original care these people are con artist reading the comments is scam when trying to update them will say that is an aftermarket phone (182)	❍	❍
my cell phone is not complete the headphones are missing they were not in the box the instructions are not the original it s a simple copy it s not as seller show in the picture i feel scammed (183)	❍	❍
after a week of using it every time i turn it on it freezes after sec i tried to reset but didnt work my phone is broken it didnt fell and i cannot use it what should i do garbage phone what should i do (184)	❍	❍
the product isn t the one as described actually the phone is the model at t sgh i instead of the international version i as described they are quite different starting with different cpu language possibles to work with g vs g ando so on (185)	❍	❍
it was great as expected i was concerned because i had read	❍	❍

many reviews saying that the telephones were from a t but this one was factory unlocked no signs no brands so far it has performed excellent coming from blackberry this is awesome (186)		
ordered this item thinking i could swap my working sim card into it this is not the case after some research i have found that this phone will not work in the us even though it is an international phone that is factory unlocked i am very dissapointed (187)	❍	❍
i love the big screen nothing i can say bad about this fon this is the best fon i ever had (188)	❍	❍
i am very disappointed because the phone is locked for my country and i have to pay a lot of money to unlock it (189)	❍	❍
it was a greaat desl wud buy one more but it took long to come but that was ok thou its nice to have a galaxy now (190)	❍	❍
this cell phone was for my daughter that lives in costa rica and it is working wonderful my daughter don t have any complaints (191)	❍	❍

sensasional phone a stylish design but a little big it is a very advanced with extraordinary functions very happy with purchase (192)	❍	❍
great phone better than iphone s easy to use and very flexible tocustomize for my personal use easy transaction and fast delivery (193)	❍	❍
my sister can t stop talking about her phone has actually stopped carrying her laptop to work even with the gs out this is a great phone (194)	❍	❍
excelent celular phone the best its very fast excelent digital images a lot of free aplications download very very fast best than iphone (195)	❍	❍
the phone is not good they send me a good battery for s but the phone is a fake because it doesn t have the same place for battery like one real and now they don t answer to me (196)	❍	❍
i hate this phone so much it was a waist of money the interface is bad and my camera has sense stopped working alone with other apps like the internet your better off to spend on a rock	❍	❍

(197) great phone great camera great wifi connection to internet no matter where you are love the big screen never will buy another phone i bought the unlocked intl version and i had no problems whatsoever	❍	❍
(198) it is a perfect phone meets all expectations the tools of visualization communication games social networking the operating system is very fast the quality and speed of the camera is great i highly recommend it	❍	❍
(199) my dad bought this for himself and he loves it lovrs it loves it he cannot get enough of his gs samsung is just amazing and amazon has a great shopping and shipping experience will definitely be shopping here again	❍	❍
(200) i bought this as a gift to my wife b day got fed up with at t contracts now i m happy and can move to any carrier only thing got concern is when i open it first time saw some asian language settings other than that everything is good	❍	❍
(201) very powerful and neat machine especially given the price and screen size it s a	❍	❍

good buy you need tweak a little to optimize the ram usage if you are a heavy user also battery does not last very long like any other smart phone at this age (202)		
dont trust him this cellphone is supposed to be unlocked international but it s not i had to find another place to unlock my phone and it cost me dollars couldnt get in touch w them because could not find an email or phone number dont buy it (203)	❍	❍
the link clearly states that it is unlocked i ordered it and because i was in a hurry to travel i didn t even review the phone before i left with it now that i ve finally had time to open the phone and actually try to use it it shows up as t mobile locked (204)	❍	❍
i have serious problems with my phone sometimes i can not use it because the touch screen is blocked and i also have problems when i want to update the system it does not work anyway i bought it when i was in the states two months ago now i can not do anything i am in romania (205)	❍	❍
finally the phone arrived today	❍	❍

in a box with a broken seal it said it was a phone for use in europe only there is no way to contact the seller to ask if there is a phone that can be used in south america my only recourse was to contact amazon and they authorized a return what a disappointment (206)		