ACCURACY AND INTERPRETABILITY TESTING OF TEXT MINING METHODS

Triss A. Ashton, B.S., MBA

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2013

APPROVED:

Nicholas Evangelopoulos, Committee Chair
Robert Pavur, Committee Member
Victor Prybutok, Committee Member
Audhesh Paswan, Committee Member
Mary C. Jones., Chair of the Department of Information Technology and Decision Sciences
Finley Graves, Dean of the College of Business
Mark Wardell, Dean of the Toulouse Graduate School

Ashton, Triss A. *Accuracy and interpretability testing of text mining methods.* Doctor of Philosophy (Management Science), August 2013, 128 pp., 51 tables, 19 figures, reference list, 100 titles.

Extracting meaningful information from large collections of text data is problematic because of the sheer size of the database. However, automated analytic methods capable of processing such data have emerged. These methods, collectively called *text mining* first began to appear in 1988. A number of additional text mining methods quickly developed in independent research silos with each based on unique mathematical algorithms. How good each of these methods are at analyzing text is unclear. Method development typically evolves from some research silo centric requirement with the success of the method measured by a custom requirement-based metric. Results of the new method are then compared to another method that was similarly developed.

The proposed research introduces an experimentally designed testing method to text mining that eliminates research silo bias and simultaneously evaluates methods from all of the major context-region text mining method families. The proposed research method follows a random block factorial design with two treatments consisting of three and five levels (RBF-35) with repeated measures.

Contribution of the research is threefold. First, the users perceived a difference in the effectiveness of the various methods. Second, while still not clear, there are characteristics with in the text collection that affect the algorithms ability to extract meaningful results. Third, this research develops an experimental design process for testing the algorithms that is adaptable into other areas of software development and algorithm testing. This design eliminates the bias based practices historically employed by algorithm developers.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Development of Big Data

Since the computer and communications technology revolution of the late twentieth century, companies have been collecting greater volumes of information about customers and products. The volume of data now held has led to a new descriptive term *big data* to refer to the massive databases held by firms. In 2011, 37% of respondents to a survey initiated by The Data Warehouse Institute (TDWI) reported holding between 10 and 100 terabits of data just for analytics (Russom, 2011). In the same report, 5% of respondents reported holding over 500 terabits for analytics while 20% of respondents expected to hold over 500 terabits just for analytics by 2014 (Russom, 2011). Most of that data is in a quantitative form; however, a greater volume of data is now in a written natural language format that is routinely called unstructured data. In the TDWI survey, 35% of respondents reported gains in the volume of unstructured data on hand (Russom, 2011).

Internal data, as described in the TDWI survey above, is not the sole source of unstructured text data available to a business. Recently social media platforms, e.g. Twitter, began emerging. These platforms allow customers and potential customers the opportunity to communicate, promote, and share information in a written format with others about any topic they desire. Those communications, referred to as tweets in the instance of Twitter, frequently include important information about businesses and products. From a business intelligence (BI) perspective, access to social media data is a valuable data source about one's own firm and products as well as competitors. This data is available at no cost in many instances, as many of the social media platforms, including Twitter, provide application programming interfaces (API)

which allow businesses to access at least some of that data by queries. Thirty four percent of firms recently surveyed reported they are now "aggressively exploring social media data" (Russom, 2011).

Collectively, this data provides an unprecedented look into customer opinion about the firm, customer opinion about the firm's competition, the needs of the customer, and opinion about products. It also represents the voice of the customer in its most powerful form available – the customers own words. Results from analyzing this data have the potential to significantly change the way business responds to the market (Jaspersoft, 2011).

Introduction to Text Mining

Analysis of text data typically involves some form of *text mining*. Text mining is an umbrella term that describes technologies used to analyze semi-structured and unstructured textual data (Miner et al., 2012, p. 30). The term is frequently used interchangeably with *text analytics* and the technologies that fall under this umbrella focus on converting text to a structured numerical representation (Miner et al., 2012, p. 30). As illustrated by the ovals in Figure 1, text mining is an interdisciplinary endeavor that involves elements of data mining, artificial intelligence and machine learning, computational linguistics, library and information sciences, databases, and statistics (Miner et al., 2012, p. 31). Text mining consists of seven distinct practice areas. These include document classification, document clustering, information retrieval, web mining, concept extraction, natural language processing, and information extraction (Miner et al., 2012, p.31). These practice areas are also depicted in Figure 1 and are plotted on their discipline of origin. This dissertation concentrates on text mining techniques that analyze a corpus for concept extraction, and then clusters documents of the corpus about those

concepts. More specifically, the interest is in the extraction of high-level latent concepts that are repeated across many documents of a corpus.

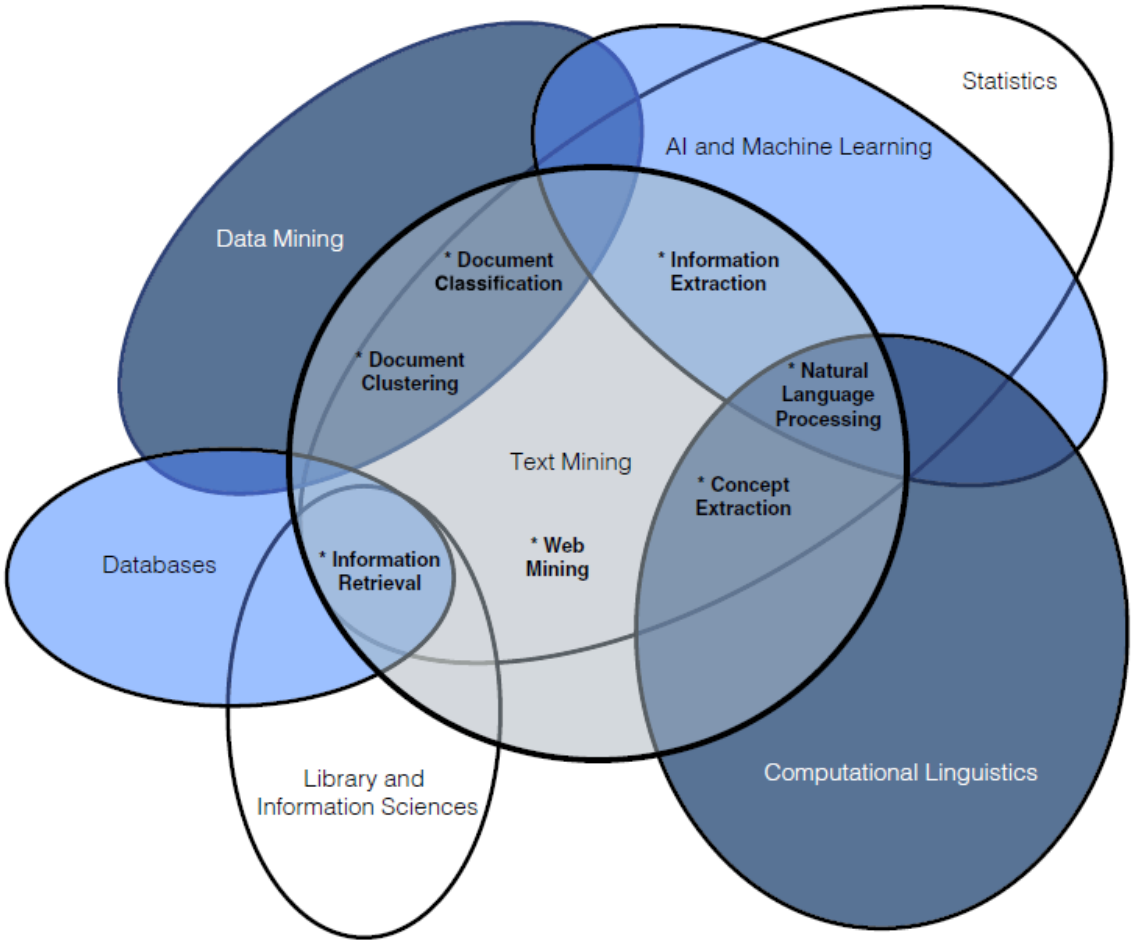

Figure 1. Venn diagram of text mining, its related fields, and key practices (Miner et al., 2012, p. 31).

Several software packages with text mining capability that perform concept extraction and then document clustering are immediately deployable by an interested organization. These packages are based on a variety of algorithms and have different strengths and weaknesses making some packages more useful than others depending on the nature of the corpus under

consideration. Determining algorithm perceived performance characteristics is the central theme of this dissertation. However, to avoid the inherent issues associated with a particular software package and its programming language, all analytic activity for this dissertation occurred within the R statistical programming language. This language is selected because it provides a wide range of programming libraries. These libraries facilitate all of the analytic procedures necessary to perform text mining across an extremely diverse mathematical range.

Broadly speaking, text mining methods are classified as either context word models or context region models. Context word models operate based on word-by-word co-occurrences while context region models function with word-by-document occurrences in a corpus. This research focuses on concept extraction as performed by context region models because they are more prevalent in business.

The context region family of models includes four subfamilies: latent semantic analysis (LSA), non-negative matrix factorization (NMF), probabilistic latent semantic indexing (pLSI), and topic models. LSA was originally introduced as a method of indexing electronic text data in information retrieval research (Dumais et al., 1988; Deerwester et al., 1990). LSA was quickly recognized as a new theory of learning, memory, and knowledge in the cognitive sciences discipline (Landauer and Dumais, 1997). The second family of text mining tools is NMF (Lee and Seung, 1999). In its introduction, NMF was illustrated as a text-mining tool; however, it has had even greater success in the hard sciences performing tasks such as DNA analysis and computational biology (Devarajan 2008). The third family of context region text mining methods, pLSI, was the first attempt at defining text in probabilistic statistical terms (Hoffmann, 1999). pLSI however, does not fully define the process statistically. That led to the development of latent Dirichlet allocation (LDA), the founding method of the topic models

family. LDA attempts to model a corpus probabilistically (Blei and Jordan, 2003). pLSI can be thought of as a bridging algorithm to LDA. Further, while pLSI is a distinct algorithm, it and NMF discover the sub-topics of the corpus by performing an optimization of the same objective function with nearly the same update rules (Ding, Li, and Peng, 2006 and 2008; Gaussier and Goutte, 2005). As a result, pLSI is not further considered in this research.

Text mining begins with a text collection called the corpus. The corpus consists of documents, which are individual communication expressions that can range from as little as a singular customer comment on some topic to a full financial report. In the context region models family, all of the text mining methods operate on a matrix-based representation of the corpus. The matrix assigns each document a unique column and terms or words found across the corpus are assigned separate rows. Cells lists the frequencies of occurrence of term *i* in document *j*. After matrix generation, the selected text mining method discovers the latent topic structure of the corpus using its inherent mathematical algorithm. As illustrated in Figure 2, after discovering the latent topics, the text mining method generates lists of key terms that define each discovered latent topic and lists of documents that are associated with each of the topics.

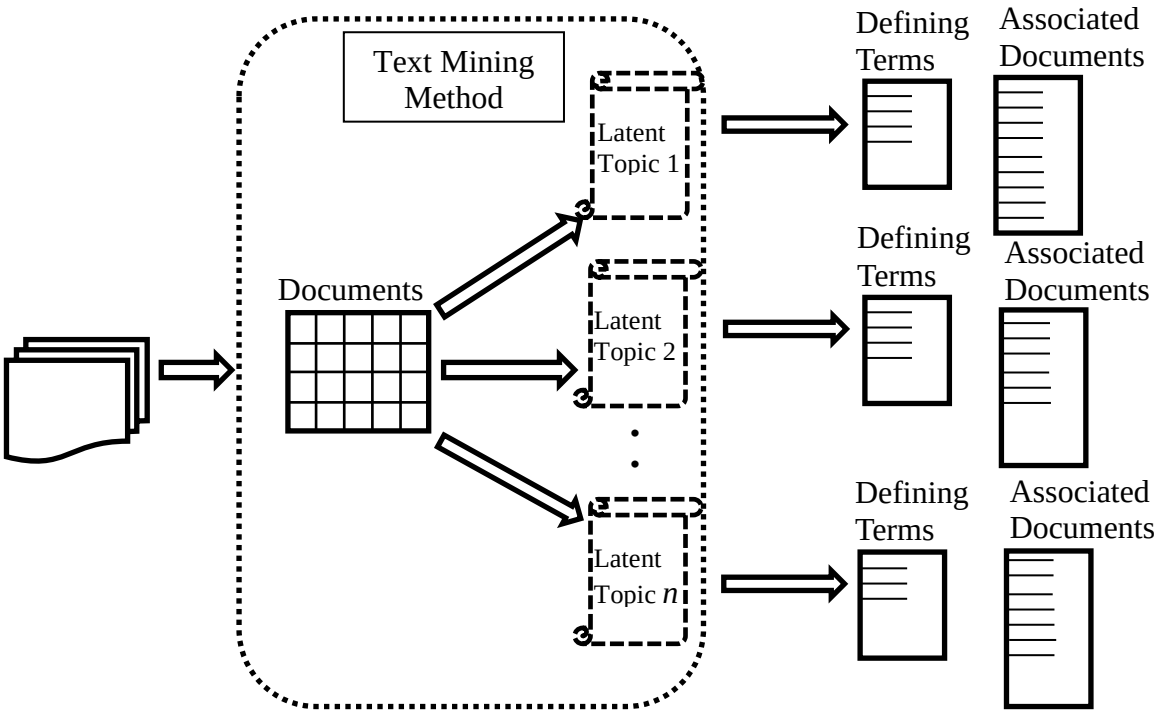

Figure 2. Context region modeling flowchart.

Past Algorithm Testing

Which text mining algorithm is best for a given situation remains an open research topic. Some algorithm testing exists in the literature but with mixed conclusions (see Table 1). For example, in three separate evaluations, LSA out performed pLSA at automatic essay grading (Kakkonen et al., 2005; Kakkonen et al., 2006; Kakkonen et al., 2008). However, pLSA outperformed LSA in two tests involving four medium sized standard document collections (Hofmann, 1999a; Hofmann, 1999b; Hofmann, 2001). pLSA also outperformed LSA at machine translation using an accuracy ratio metric (Kim et al., 2002). In two evaluations, LDA underperforms both LSA and pLSA at automatic essay grading (Kakkonen et al, 2006; Kakkonen et al., 2008). However, LDA outperformed pLSA in precision particularly when the number of topics is high (more than 100 topics) (Chang et al., 2009). In testing conducted in

cognitive research, results for LDA and LSA alternated back and forth across five datasets and in many instances were apparently indistinguishable (Riordan and Jones, 2011). No testing exists in the literature comparing LDA, LSA, and NMF in a single setting.

Table 1

Past Text Mining Testing Studies

Compared Algorithms	**Text Type or Task Performed**	**Measurement Metric**	**Most Efficient Algorithm**	**Study**
LSA vs. pLSA	Medium length abstracts (MED, CRAN, CACM, CISI)	Precision Recall	pLSA	Hofmann, 1999a
LSA vs. pLSA	Medium length abstracts (MED and LOB	Perplexity	pLSA	Hofmann, 2001
pLSA vs. LDA	Avery abstracts and TREC AP newswire	Perplexity	LDA	Blei et al., 2003
LSA vs. NMF	NIST's TDTC and Reuters-21578	Accuracy and Normalized Mutual Information Metric	NMF	Xu et al., 2003
LSA vs. pLSA vs. pLSA-C	Essay Grading at sentence and paragraph level	Compared to human grading	LSA	Kakkonen et al., 2005
LSA vs. pLSA vs. LDA	Automatic Essay grading	Compared to human grading	LDA<pLSA<LSA	Kakkonen et al., 2006 and 2008
pLSA vs. LDA	New York Times articles and Wikipedia articles	Model Precision	LDA* (at lower dimensions, $k \leq 100$, pLSA matched LDA	Chang et al., 2009
LSA vs. pLSA vs. LDA vs. CTM	Topic Detection and Spam Filtering. Customers comments on cameras at Amazon.com	Spam: Precision & recall Detection: by Descriptive	pLSA for Spam Filtering	Lee et al., 2010
LSA vs. LDA vs. Context Word models	McRae, 2005; Vinson & Vigliocco, 2008; MCDI 1994;	Purity and Entropy	Context Word Models	Riordan and Jones, 2011
LSA vs. LDA vs. Construction-Integration	TOEFL and Nelson Norms	Predictions	Regarding LSA and LDA – "no obvious general superiority"	Kintsch and Mangalath, 2011

This pattern of inconsistent results is prevalent throughout the text mining testing literature. Further, there appears to be a pattern where the authors of new algorithms design tests

using either data sets or measurement metrics that are optimal for the envisioned algorithm. This is an intuitively logical choice; design a method for a purpose, then test the method performing that task. Frequently the new algorithms are compared to older algorithms that were not designed for the task to begin with.

An evaluation typically involves measuring task performance e.g. document classification, or estimating probabilities for holdout data (Wallach et al., 2009). With the exception of Chang et al., (2009), all text mining testing tends to follow one or more of the measured task performance methods described by Wallach (2009). However, Wallach et al. (2009) found that these evaluation methods are inaccurate. The patterns observed in past testing results suggests a need for an experimentally driven test method that looks at the text mining results comprehensively and guards against bias.

An alternative to measuring task performance is to have potential user evaluate the analytic results generated by the text mining methods. The first attempt at this approach measured human evaluator's perceptions of the results using semantic coherence and association coherence constructs (Chang et al., 2009). Semantic coherence measures how well the terms belong together and form topics. More specifically, Chang et al. (2009) used semantic coherence to measure word intrusion, which is the presence of seemingly unrelated words to a list of topic defining terms. Semantic coherence of terms, however, is only valid if the topic structure is horizontal and not hierarchical in nature. A corpus with a horizontal structure consists of documents on topics that are unrelated or marginally related; a corpus with a horizontal topic structure is heterogeneous in nature. A hierarchical topic structure, a homogeneous corpus, possesses interconnected topics that are a mixture of broad topics, subtopics, and sub-subtopics.

In a homogeneous corpus, individual communications reflect related topics in the hierarchy because of word select. This makes the semantic coherence of terms inaccurate in some corpora.

Association coherence measures how well the algorithm classified documents to the topics. Semantic coherence then measures internal consistency of the terms and association coherence measures construct validity of the documents. This application of human evaluation did not address the internal consistency of the documents or the construct validity of the terms.

A more universal approach begins by recognizing that text mining methods perform two fundamental functions: identify latent topic structures in the corpus and relate documents to those topics. Testing the quality of the results with potential users would measure how understandable the extracted topic is, how well the defining terms describe the topic, and how well the associated documents reflect the topic.

Research Design, Question, and Contribution

This research considers the text mining methods LDA, LSA, and NMF focusing on the clarity or understandability of the results generated by the analytic method. Evaluation of clarity requires the measuring of the linkage between the latent topics and the associated documents, and the link between latent topics and the defining terms. To begin to address these issues, a homogeneous corpus with an unknown internal structure was acquired. A homogeneous corpus is desirable because in a business context, most corpora are homogeneous as a result of internal data management practices. Discovery of the latent structure in the corpus is accomplished by the text mining methods, and then potential human users evaluate the results. The human evaluation phase looks at the algorithms ability to extract a distinct topic and then associate

document and terms to that topic that are meaningful to humans. The research question for this experiment is:

RQ: Which algorithms do humans perceive as more effective at extracting topics and classifying documents to topics?

This research is very important to the business and business research community. Thirty-three percent of firms worldwide indicated they use data mining, the analytic family to which text mining belongs, to help them understand the customer (Allen, Gearan, and Rexer, 2011). Text data found in unstructured databases is the voice of the customer and it is in the customers own words. However, before the decision-making process can integrate text-mining results, the accuracy of the algorithm must be understood. Text-mining methods have developed incrementally. Evaluation of existing methods reveals strengths and weaknesses and is important for guiding future developments as well as the refinement of existing algorithms.

CHAPTER 2

TEXT MINING ALGORITHMS

Composition of a Corpus

The collection of raw data that text mining is applied to is called a corpus. The corpus consists of textual entries composed in natural written language. In the business framework, each uniquely composed entry is managed separately and the entries are called documents even though some consist of little more than 10 word comments. The collection can be composed by one or many writers. The structure of the discussion topics in a corpus can be heterogeneous with a variety of generally unrelated subtopics; homogeneous with the subtopics related to one or more broad topics; or a mixture of the two. Businesses typically manage document collects along the organizational structure, which by default leads to a topic structured database such that customer composed documents are housed separately from internally generated documents. Further, businesses tend to partition customer documents even further such that customer service related complaints are housed in a database separate or are otherwise distinguishable from product satisfaction statements. As a result, business corpora are either homogeneous or a mixture in nature.

Within a corpus, documents can consist of a collection of very short text expressions e.g. Twitter postings, which are limited to 140 characters, or much larger textual items e.g. research reports, which can consist of thousands of words. Because business corpora are managed by topics, they also tend to consist of documents that are similar in length. Many would dismiss this point, however, all text mining methods function on the Term × Document matrix. The Term × Document matrix holds frequency information and therefore reflects a probability distribution function (pdf). As average word counts in documents change, the distributional characteristics

of the corpus should change. The nature of that change is currently unknown and could consist of simple shifts in skewness and kurtosis or could shift the distribution to a completely different pdf.

Historical Developments of Text Mining Algorithms

Text mining is an extension of information retrieval methods. Information retrieval is defined as simply finding material, usually documents of an unstructured nature from a large collection that satisfies a need (Manning et al., 2009, p. 1). Information retrieval started by a machine reading text looking for matches to retrieval terms (Manning et al., 2009, p. 3). This very time-consuming process led to indexing of documents (Manning et al., 2009, p. 3). Indexing led to the discovery of term-document incidence matrices and searches that were more complex. A term-document incidence matrix cells are filled with 0's and 1's that indicate the term is or is not present in a certain document. The resulting matrix is extremely sparse with only 1-to-2% of the cells consisting of a one (Manning et al., 2009, p. 6). With the term-document matrix, a Boolean retrieval that mixes the operators AND, OR, and NOT, for example, *term A*, and *term B*, and not *term C* is possible (Manning et al., 2009, p. 4). However, as the amount of text in the database expands so does the size of the term-document matrix and eventually it is unable to fit in the computer's memory. The sparsity condition of the matrix and memory limits led to the development of the inverted index. The inverted index consists of a dictionary terms list and a list that records which documents the term occurs in (Manning et al., 2009, p. 6).

Throughout this early development period, it was learned that retrievals based on raw frequency are problematic. Many of the index-based searches tend to return documents based on how many times the search term appeared in it. Second, with regards to the corpus as a whole,

high frequency terms do not necessarily convey meaningful information. Prepositions and conjugates, for example, are by far the most frequent terms in a document, however they convey almost no meaningful information; instead, they fill in the space around meaningful words and provide connection among ideas. High frequency words that only appear in select documents, however, are important to defining topics.

These insights led to the development of the vector space model (VSM) (Salton et al., 1975). VSM is fundamental to many information retrieval operations including scoring documents on a query, document classification, and document clustering (Manning, 2009, p. 120). The VSM model concluded that the best index models had more vector space between terms and introduced term weighting as a method of elevating terms that occurred with high frequency in only individual documents while simultaneously remaining rare in the corpus (Salton et al., 1975). Term weighting then shifts the natural probability of the Term × Document matrix.

Overview of Modern Text Mining Algorithms

The modern text mining methods extended from these early developments progressing in a variety of directions. The first major group to emerge was the context region model, which perform analysis on a term-by-documents frequency matrix. This group of models focused on latent structures, which are described as topics, concepts, or dimensions, and began being introduced in 1988. More recently, a second group of text mining methods called the context word models emerged and are based on a word-by-word frequency matrix (Riordan and Jones, 2011). The word-by-word frequency matrix of a context word model holds the frequencies of words-by-word co-occurrences that appear within the range of a moving window (Riordan and

Jones, 2011). To date, the contest word model appears to have a stronger orientation to cognitive research. There are no active research lines I am aware of relating context word models to business corpora. In this dissertation, I focus only on the context region model text mining methods that quantify language usage patterns in the form of latent topics, concepts, or dimensions. The end-products of these methods are groups or clusters of terms and documents that are associated with the latent topics, concepts, or dimensions.

Statistically speaking, the modern context region class of text-mining methods can be broadly classified as descriptive or probabilistic. The modern descriptive text-mining algorithms include LSA and NMF families. The modern probabilistic text-mining methods include pLSI, and the topic models families of algorithms.

Text mining algorithms have been evolutionary in their development. The modern methods build one upon another and descend from the vector space model (VSM) algorithm. VSM is an algebraic method of representing documents as vectors that originated as a tool for information retrieval and indexing (Sultan et al., 1975). However, VSM lacks semantic sensitivity, that is, retrieval based on conceptual content was not possible. Further VSM does not handle synonyms (words with similar meanings) or polysemy's (words with multiple meanings) very well (Deerwester et al., 1990). These weaknesses led to the development of LSA.

The original conceptualization of LSA was to reduce the dimensionality of information retrieval problems (Dumais et al., 1988). LSA assumes an underlying latent structure exists in word selection that is obscured by word selection variability (Dumais et al., 1988). LSA was motivated as a method of retrieving information based on conceptual content instead of just individual words (Deerwester et al., 1990). The primary mathematics behind LSA is an operation of linear algebra called singular value decomposition (SVD). LSA captures the

similarities of words so well that with each new paragraph of text it encounters, the model improves at a rate that approximates the natural learning rate of young schoolchildren (Landauer and Dumais, 1997). The LSA model "exhibits humanlike generalizations" and can be used to infer indirect similarity of meaning (Landauer and Dumais, 1997). Psychologically speaking, LSA has been described as a theory of learning, memory, and knowledge analogous to a neural net model (Landauer and Dumais, 1997). The LSA model has also been used to distinguish the writings of different authors (Nakov, Popova, and Mateev, 2001).

NMF is defined as "… a method for modeling the generation of directly observable visible variables *V* from hidden variables *H*" (Lee and Seung, 1999). NMF estimates an input matrix as two variable matrices by optimization and has a stochastic nature (Brunet et al., 2004). As its name implies, negative values are not allowed in the factor matrices, those matrices on the right hind side of the equality. The non-negative constraint compliments the intuitive notion of a whole formed by its many parts (Lee and Seung, 1999). NMF actually extends from a line of research called positive matrix factorization (PMF). PMF argued that in the physical sciences the origin was well-defined and that centering the data, as is the case in principal components, caused a loss of information (Paatero and Tapper, 1994).

In its introduction, NMF was demonstrated in image analysis and semantic analysis (Lee and Seung, 1999). In recent years, many variations of the NMF algorithm have been introduced (Pascyal-Montano et al., 2006; Gaujoux and Seoighe, 2010). These NMF algorithm variations typically alter the optimization update expressions or the minimization expressions. NMF has been applied effectively in many domains of science that involve sparse matrices including computational biology (Devarajan, 2008), gene expressions (Badea, 2008), low-resolution brain electromagnetic tomography (Pascual-Montano et al., 2006), metagenes and molecular pattern

analysis (Brunet et al., 2004), and text mining (Lee and Seung, 1999; Pauca et al., 2004; Shahnaz et al., 2006).

While pLSI is not considered in this research, it is briefly mentioned here because of its contribution to subsequent models. pLSI, also referred to as probabilistic latent semantic analysis (pLSA) was an important first step at automated document classification along a probabilistic statistical model (Hoffmann, 1999a and 1999b). pLSI is based on the likelihood principle and its core model is referred to as the *aspect model* (Hofmann, 1999a). pLSI considers documents a mixture with each word a sample from separate topics and each of the topics distributed as multinomial random variables (Blei, Ng, and Jordan, 2003).

pLSI is considered statistically incomplete because it does not model the documents probabilistically (Blei, Ng, and Jordan, 2003). This led to the introduction of a new family of text mining tools, now broadly referred to as *probabilistic topic models* or more commonly just *topic models*. Latent Dirichlet allocation (LDA) was the first of the topic models methods. LDA represents documents as a random mixture of topics and it considers topics as distributed over words (Blei, Ng, and Jordan, 2003). A document can consist of one or many topics and a topic is distributed over a fixed vocabulary (Blei, 2011). In LDA, topics are considered distributed as a Dirichlet distribution. For parameter estimation, LDA implements a variational estimation maximization (VEM) algorithm. In the paragraph that follows, LDA is referred to as LDA-VEM to distinguish between the original LDA and the many variants that followed.

Following the introduction of LDA-VEM, many variants were introduced with each oriented toward either filling a specific need or moderating a perceived problem with the LDA-VEM algorithm. These variants are important to the development of the topic model family and play a role in selecting algorithms. Therefore, they are worthy of mention with a brief note on

their contribution. The first variant introduced was hierarchical LDA (hLDA) (Blei et al., 2004). hLDA assumes a corpora consists of topics that are framed around a hierarchical structure and introduced the concept of Gibbs sampling to topic models (Blei et al., 2004). Dynamic topic models (DTM) considers topics as evolutionary over time and applies Kalman filters and non-parametric wavelet regression to the approximation of posterior probabilities (Blei and Lafferty, 2006). Supervised LDA (sLDA) adds a response variable to LDA which is associated with each document and is then able to predict some outcome, e.g. movie ratings based on the number of stars (Blei and McAuliffe, 2007). Gibbs sampled LDA (GibbsLDA) uses a Markov Chain Monte Carlo algorithm for inference instead of the variational expectation maximization of the original LDA (Griffiths and Steyvers, 2004; Wei and Croft, 2006; Phan et al., 2008; Porteous et al., 2008; Chang, 2010). GibbsLDA requires less memory and speeds up analysis (Griffiths and Steyvers, 2004; Porteous et al., 2008). Correlated topic model (CTM) is a variant of LDA-VEM that assumes a correlation among topics (Blei and Lafferty, 2007). To accommodate the correlation, CTM replaces the Dirichlet distribution with a logistic normal distribution (Blei and Lafferty, 2007). However, CTM only captures correlation between pairs of topics (Li and McCallum, 2006). This led to the Pachinko allocation model (PAM) which is capable of capturing arbitrary, nested, and sparse correlations (Li and McCallum, 2006).

In this research, the algorithms of interest are LSA, NMF, and the original LDA model herein referred to simply as LDA. LSA is selected because of it is the cornerstone to the modern text mining tools and it is very important in business research. Of the top five most frequently used commercial text mining packages, LSA is the underlying technology for the 1st (Statistica), 3rd (SAS), and 5th (RapidMiner) packages (Allen, Gearan, and Rexer, 2010). NMF is selected because only limited testing of the LSA versus NMF algorithms exist in the literature and to

date, no testing of NMF versus topic models has been conducted. LDA is selected from the topic models family because it is the principal algorithm. Many of the variants that followed are more-or-less task specific applications. Finally, there are no results in the literature involving the simultaneous testing of these three algorithms.

Common Processing

All of the modern text-mining methods share a number of common data processing steps in preparing the term-document matrix. Data preparation begins with a series of preprocessing steps applied against the individual documents of the corpus. Typically preprocessing starts with the removal of punctuation and in most cases, removal of Arabic numbers. Low information words, commonly called stop-words, are also removed from the data. Stop-words are terms that exist in the corpus with high frequency and dominate the probability distribution yet they convey little information of value.

Stop-words are commonly combined into tailored lists and include, for example, the conjunctions (*and, but, or, nor, for, so,* or *yet*), prepositions (*about, above, at, beyond, until,* or *with*) and interjections (*ouch, oh no,* and *hey*). Stop-words routinely have additional words included that occur with extremely high frequency across the corpus. For example, the word *restaurant* would occur with very high frequency in a dataset of customer comments held by a food service firm. In this instance, *restaurant* would be understood; however, its high frequency will exert pressure on the outcome beyond the value it adds.

Finally, the words are stemmed or truncated back to their roots. For grammatical effect, words use a variety of prefixes and suffixes. Text mining is not concerned with different tense forms of a word. A discussion topic occurring in the past, present, or future is fundamentally the

same discussion topic. By stemming, words with a common root are counted as a single entry in the term-document matrix. In information retrieval, the Porter stemmer (Porter, 1980) or the Snowball stemmer (Porter, 2001) are commonly used to perform stemming.

Following corpus preprocessed, the term-document matrix $X_{t\times d}$, a $t\times d$ frequency count matrix with t terms and d documents is extracted from the corpus. The matrix is evaluated along the terms and documents. This evaluation ensures words occur with some minimal frequency by summing the term row and that the word occur in some minimal number of documents by summing the number of columns with cell values greater than zero. At this stage, the $X_{t\times d}$ matrix is a term-by-documents matrix containing the raw frequencies of each term in each document and it is ready for application of algorithm specific analysis.

Latent Semantic Analysis

LSA begins by further processing the raw frequency X matrix into an equivalent of the vector space model (Salton, et al., 1975). This process involved the transformation of the count data found in the X-matrix by a weighting method. Term weighting conditions the data and involves a local weight and a global weight. Similar to adding high frequency words to a stopword list, weighting addresses terms that appear with high frequency across the document collection. For example, the inverse document frequency (idf) method reduces the impact of terms that appear across the collection with high frequency and instead favors those high frequency terms that appear in relatively fewer documents (Salton and Buckley, 1988).

Weighting is achieved by taking the frequency of each cell of the X-matrix and adjusting it by taking the product of a local and a global weight for that cell. Local weights transform the frequency of term i in document j (tf_{ij}) into a relative frequency. Several local weighting options

are available. The most common term frequency weighting method involves simply using the observed tf_{ij} value as the local weight. Other alternatives for local weighting include binary where the local tf_{ij} value is taken as either 0 (does not exist) or 1 (exists). Log local weighting replaces the tf_{ij} value with $\log(tf_{ij}+1)$) (Salton, Buckley, 1988). Finally, augmented normal (augnorm) (replaces the tf_{ij} value with $0.5+0.5*tf_{ij}/\max(tf_{ij})$) (Salton, Buckley, 1988).

Global weights describe the relative frequency of the term across the corpus. Several global term weighting methods are available including idf, entropy, and binary. Idf is defined as $\log_2(N/n_i)$, where N is the number of documents in the collection and n_i is the frequency of term i in the corpus. Entropy global weighting is defined as $1+[\Sigma p_{ij}*\log(p_{ij})/\log(N)]$ where $p_{ij} = tf_{ij}/n_i$ is the conditional probability for the document j given term i.

The proper selection of a term weighting method appears to relate to the length of the documents in the corpus. The length of the document drives the sparsity of the input matrix which in-turn influences the probability of a document in a corpus and the probability of a words relating to a topic. The two highly recommended and most commonly cited weighting methods in the LSA literature are tf-idf and log-entropy (Evangelopoulos et al., 2012). Generally tf-idf is recommended when the corpus consists of large complex term structures and log-entropy is recommended when the corpus consists of smaller latent structures composed of a few frequent terms (Evangelopoulos et al., 2012).

Another common procedure is to normalize the X-matrix once it has been weighted. Normalizing the matrix equalizes the lengths of widely varying vectors. When a corpus consists of a mixture of short and long documents, the long documents exert greater influence. In information retrieval, long length documents were found to have a greater probability of being returned than short document. It should be obvious that there is a greater probability of finding a

random term in a long document than a short document. Normalizing allows all documents to be treated equally relevant (Salton, Buckley, 1988). To normalize an X-matrix, let *w* represent the weighted term value, then the normalized value is $w/\sqrt{[\Sigma(w_i)^2]}$ (Salton, Buckley, 1988). With the X-matrix normalized, it is ready for decomposition.

The primary analysis procedure of LSA is singular value decomposition (SVD). SVD is a linear algebra technique that decomposes any rectangular matrix into three other matrices

$$\mathrm{X}_{t\times d} = \mathrm{U}_{t\times r}\mathrm{S}_{r\times r}\mathrm{V}^{T}_{r\times d}\,. \qquad (1)$$

The $\mathrm{U}_{t\times r}$ matrix is terms-by-factors in dimensionality and is the eigenvectors of the XX^{T} matrix. The $\mathrm{XX}^{\mathbf{T}}$ matrix is a $t\times t$ dimension term covariance matrix that define *r* latent semantic themes in the data that are called factors in the multivariate language. The terms of each factor define the latent semantic topics. $\mathrm{S}_{r\times r}$ is a $r\times r$ diagonal matrix of singular values. $\mathrm{V}^{\mathrm{T}}{}_{r\times d}$ is a factor-by-document matrix that represents the eigenvectors of the $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ matrix, a $d\times d$ document covariance matrix which associates factors and the original documents.

For a variety of reasons, it is desirable to reduce the number of factors extracted from the X-matrix. In a population, removing those factors with eigenvalues less than one places a lower bound on the number of common factors (Guttman, 1954). When a factor accounts for less variance than a single variable, it is of little interest (Cliff, 1988). Frequently, the data are in high-dimensionality space while only a few dimensions convey the topic structure of the corpus (Zhu and Ghodsi, 2006). The SVD products are truncated to a reduced space of only the first *k* highest rank singular values, $s_1, \ldots, s_k$, such that $\widehat{\mathrm{X}}_{t\times d} = \mathrm{U}_{t\times k}\mathrm{S}_{k\times k}\mathrm{V}^{T}_{k\times d}$ is the best *k* rank least squares estimate of $\mathrm{X}_{t\times d}$. What *k* value constitutes the selection of the best *k* singular values is still an open research topic. For further discussion on the best *k* singular values to extract, see the section *Determining the Number of Topics to Extract* later in this chapter.

Once the data are truncated, the two loading matrices, L_T and L_D, are recovered. Using the orthonormality property $V^TV = I$, where I is the $k \times k$ identity matrix, we obtain

$$L_T = U_{t\times k}S_{k\times k} = \widehat{X}_{t\times d}V_{d\times k}. \quad (2)$$

The matrix L_T consists of term loadings and associates the various terms with specific topics. It is a $t \times k$ dimension matrix. The term-factor relationship found in this matrix facilitates topic labeling. To obtain document factor loadings, we use the orthonormality property $U^TU = I$, where I is the $k \times k$ identity matrix, and $S=S^T$. Post-multiplying both sides by $U_{t\times k}$ gives

$$L_D = V_{d\times k}S_{k\times k} = \hat{X}^T_{d\times t}U_{t\times k}. \quad (3)$$

The L_D matrix is a $d \times k$ matrix of document loadings that associates the various document with specific topics.

Factor rotation has long been known to simplify a factor structure and theoretically achieves a more meaningful solution (Hair et al., 2006, p. 123). Factor rotation has also been show to improve interpretability of LSA results (Sidorova et al., 2008). Many different methods of factor rotation exist (Kim and Mueller, 1978, p. 29-40). Most of these methods have not been tested in text mining, however, Varimax rotation has been used successfully (Sidorova et al., 2008). Varimax rotation maximizes the sum of variance for the squared loadings. Rotation can begin with either the term loadings L_T matrix or the documents loadings L_D. Beginning with the L_T matrix is the recommended strategy because it facilitates factor interpretation (Sidorova et al., 2008). Once a solution matrix M is recovered, it is also applied to the L_D matrix (Sidorova et al., 2008).

Factors represent discussion topics. These topics are defined by the associated words found in the L_T matrix. The topics are discussed in the documents listed in the L_D matrix. As one moves down the list of documents associated with any given topic, loading values become

small. Eventually, documents will be associated with the topic *k* only by chance because of coincidental cross usage of one, two or three terms that define topic *k.* This is considered noise. To avoid mixing noise into an analysis, the practice of suppressing results, replacing the loading with zero, below some threshold has been adopted from multivariate analysis. Suppression levels are set at a value that represents the point where the number of loadings retained in the L_D matrix equals the number of documents in the corpus. This strategy means some documents may not be associated with any topic while others are associated with more than one topic.

Once factors are rotated and loadings are suppressed, results are ready for interpretation or follow-on analysis by other means that satisfy the researcher's objectives. An important yet typically ignored observation about the *k* extraction process in LSA is that the information to the right of selected k^{th} factor is simply deleted. If a corpus has a hierarchical structure to its topics, then those very specific sub topic toward the bottom the hierarchical tree should end up represented by those *k* values in the outer right tail of the **S** matrix. This means that if multiple branches exist in the topic structure, those topics with longer or more diverse sub structures could get truncated and not fully appreciated in subsequent analysis.

Latent Dirichlet Allocation

Statistically speaking, LDA is a much more rigorous theory than LSA. LDA models a corpus on a probabilistic basis. LDA assumes documents *D* consist of *N* words that discuss one or more individual topics *Z* and that the topics discussed are a random mixture of latent topics θ (Blei, Ng, and Jordan, 2003). Topics follow a multinomial distribution of $p(z_n \mid \theta)$. Each topic *Z* is distributed over words *w* with some probability β (Blei, Ng, and Jordan, 2003). The number of words *N* in a document *D* follow a Poisson distribution which is independent of all other

variables and considered an ancillary variable that is ignored. For a corpus, the *k* topics are defined by a word vocabulary *V* that follows a word probabilities distribution that is parameterized with $\beta_{k \times V}$ a $k \times V$ matrix that contains each word probability $\beta_{ij} = p(w^j = 1|z^i = 1)$ (Blei, Ng, and Jordan, 2003). Word probabilities β matrix is considered a fixed quantity that is estimated. Given a topic z_n, choice of the n^{th} word w_n follows a multinomial distribution with probability $p(w_n|z_n,\beta)$. Topic mixtures follow a multinomial distribution for the latent topic mixture θ and has the probability density function

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ \ldots \ \theta_k^{\alpha_k - 1} \tag{4}$$

where each of the parameters α_i theoretically can exist independently on the interval $0 < \alpha_i < \infty$ creating an infinite number of possible outcomes. To simplify this, a constant α parameter is used such that $\alpha_i = \alpha_j = \alpha$ for any possible *i* and *j*. With α fixed, and β estimated, the joint distribution of θ, *z*, and *w* is given by

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta). \tag{5}$$

In this instance, $p(z_n \mid \theta)$ is taken as θ_i such that $z_n^i = 1$. A graphical representation of the LDA model is provided in Figure 3. The outer plate is represents a document *M* from the corpus that contains a mixture of topic. Each topic is defined by word choice with some probability.

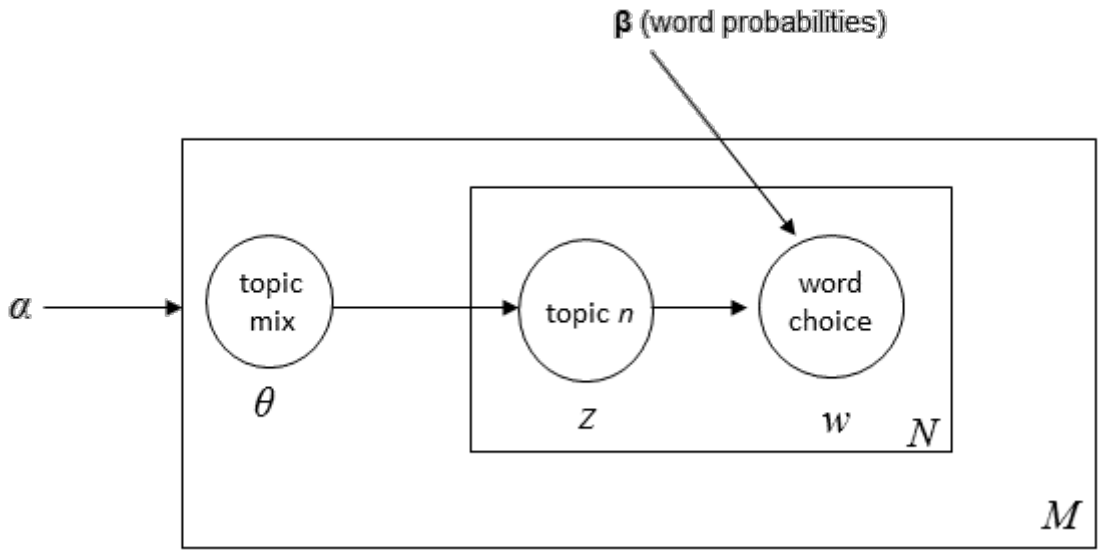

Figure 3. Graphical model of LDA.

Inference and parameter estimation for LDA, however, becomes a problem. It starts with the posterior distribution of the hidden variables

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta,z,w|\alpha,\beta)}{p(w|\alpha,\beta)}. \quad (6)$$

Unfortunately, the LDA posterior distributions are intractable because of the edges between θ and β (see Figure 3). An alternative method is to drop w and insert free variational parameters as illustrated in Figure 4. This reduces the problem to a variational distribution

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n), \quad (7)$$

with Dirichlet parameter γ and multinomial parameters ϕ are free variational parameters (Blei, Ng, and Jordan, 2003).

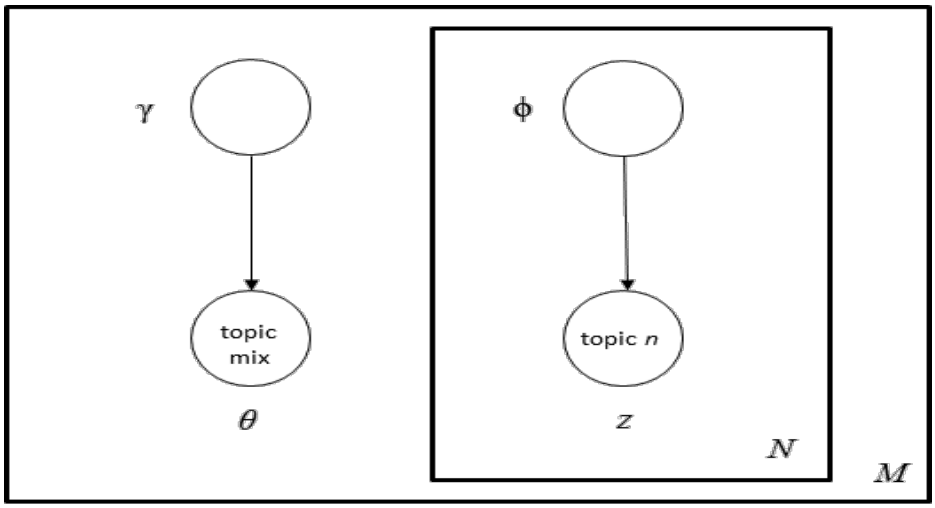

Figure 4. Variational distribution model.

Parameters are estimated in a two-step process. First (E-step), the variational parameters γ and ϕ are recovered by finding the tightest possible lower bounds with an optimization of

$$(\gamma^*, \phi^*) = argminD_{(\gamma,\phi)}(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \quad (8)$$

which leads to an appropriate starting point for γ and ϕ_n (Blei, Ng, and Jordan, 2003). With variational parameters recovered, the next step (M-step) is the recovery of parameters α and β that maximize the marginal log likelihood.

During the second step, the update for the conditional multinomial parameter β is

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dni}^* w_{dn}^j. \quad (9)$$

The α parameter is recovered by iterating over the Newton-Raphson optimization

$$\alpha_{new} = \alpha_{old} - H(\alpha_{old})^{-1} g(\alpha_{old}). \quad (10)$$

This can cause problems in instances where words in a new document do not match any of the words in the training corpus. To avoid this, variational inference methods are applied that includes Dirichlet smoothing (Blei, Ng, and Jordan, 2003).

As an iterative process, LDA requires considerable computer power. Each iteration process of the variational inference (E-step) requires $O((N+1)k)$ operations or empirically, for a single document roughly N^2k operations (Blei, Ng, and Jordan, 2003). The R statistical package Topic Models sets the variational inference (E-step) defaults at 500 iterations with a convergence tolerance of 10^{-6}; and 1,000 iterations with a tolerance of 10^{-4} for the M-step (Grun and Hornik, 2011). Preliminary testing of small corpora with the R statistical programming packages LDA implementation found configurations with repeated random starts varying from 1-30, iterations from 1,000-100,000, and tolerance levels at 10^{-6} or lower for both the E-step and the M-step were required. These configurations were not necessary to obtain convergence, but were required to obtain desirable posterior probabilities. For example, with one dataset of less than 1,000 documents, LDA obtained convergence with the default configuration; however, the high end of the associated probabilities only reached the 24% range. By changing the random starts to 30, the posterior probabilities jumped into the 90-95% range.

Non-Negative Matrix Factorization

Like LSA, NMF extends from factor analysis. NMF began in what was called Positive matrix factorization (PMF). PMF recognized that factor analysis required data to be centered; however, the data frequently was generated in some real world environments beyond behavioral science, such as physics, and consisted of well-defined origins or zero points (Paatero and Tapper, 1994). Centering data results in a loss of information about the origin and the scale of the data (Paatero and Tapper, 1994). Paatero and Tapper (1994) also found the approach of the factor problem by means of the covariance matrix inappropriate for physical science. Paatero and Tapper (1994) also viewed the X matrix as a sum of X matrices with rank one, e.g. $X \approx X^1 +$

$X^2 + \ldots + X^p$ which is a very appropriate view for text mining where each column of the X matrix (X^p) represents a document. The PMF model is then X = GF + E, where X and E are of dimensions $n \times m$, G is $n \times p$, and F is $p \times m$.

Lee and Seung (1999) introduced the first algorithm of NMF by extending PMF's idea and adopts an iterative method using the model

$$X_{n\times m} \approx (WH)_{n\times m} = \sum_{a=1}^{r} W_{n\times a} H_{a\times m}, \tag{11}$$

iterating over the set of update rules

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{X_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \tag{12}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}} \tag{13}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{X_{i\mu}}{(WH)_{i\mu}} \tag{14}$$

until a local maximum is obtained in the objective function

$$F = \sum_{i=1}^{n} \sum_{u=1}^{m} [X_{iu} \log(WH)_{iu} - (WH)_{iu}]. \tag{15}$$

This initial algorithm was shown to be functionally adequate for text mining (Lee and Seung, 1999). However, many follow-on algorithms were quickly introduced. In preliminary evaluations, the Brunet algorithm (Brunet et al., 2004) produced results that were more consistent with LSA and LDA than did the Lee and Seung (1999) algorithm. Further, when the Brunet algorithm is operationalized in C++ programming, it is almost twice as fast as the Lee algorithm (Gaujoux, 2010).

The Brunet et al. (2004) algorithm starts like the lee algorithm with the model X ≈ WH. The algorithm begins by initiating the W and H matrices randomly. It then iteratively updates the matrices minimizing a divergence function (Brunet et al., 2004). The divergence function

$$D = \sum_{i,j} X_{i,j} \log\left(\frac{X_{i,j}}{(WH)_{i,j}}\right) - X_{i,j} + (WH)_{i,j}, \tag{16}$$

is updated using

$$\mathrm{H}_{au} \leftarrow \mathrm{H}_{au} \frac{\sum_i \mathrm{W}_{ia} \mathrm{X}_{iu} / (\mathrm{WH})_{iu}}{\sum_k \mathrm{W}_{ka}}, \tag{17}$$

and

$$\mathrm{W}_{ia} \leftarrow \mathrm{W}_{ia} \frac{\sum_u \mathrm{H}_{au} \mathrm{X}_{iu} / (\mathrm{WH})_{iu}}{\sum_v \mathrm{H}_{av}}. \tag{18}$$

The function is related to the Poisson likelihood of generating X from W and H (Brunet et al., 2004).

Because of the random number initialization of the W and H matrices, convergence to the same solution may not occur with successive runs with local minima and maxima conditions found. Therefore, repeated runs are implemented. Typically 20-100 runs are sufficient to obtain a stable result (Brunet et al., 2004).

Since NMF optimizes over W and H to obtain estimates of X, reducing the number of *k* topics extracted is not as destructive as with LSA. With NMF, subtopics are folded into the higher level topics in the hierarchy. Unfortunately, the definitions of the topics could also change as a result in the new approximation.

Determining the Number of Topics to Extract

All of the text mining methods have two common problematic open research topics: *how many topics are present in the data* and *how many topics should be extracted from the data*? Little to no research exists on the first topic, *how many topics exist in data*. Most business corpora possess a hierarchical subtopic structure that fans out from some broad general topic theme, e.g. *product quality.* With each sub layer of the hierarchy, the topics become more specific, e.g. *the customer opinion of poor product quality was driven by low observed*

durability. The question of how many topics exist in a corpus is very subjective and depends on where the researcher decides to stop delineating among subtopics.

Research in text mining has then focused on the second question; *how many topics should be extracted from the data*? The approaches used typically involve some metric, however, all involve some degree of subjectivity, and there use requires a degree of expertise.

At LSA's inception, the number of dimensions to extract extended from the already existing principal components analysis methods. Those techniques for determining *k* can be broadly classified as stopping rules based on confidence intervals, or stopping rules based on average test statistic values (Peres-Neto et al., 2005). These methods included the Bartlett's test for equality (Bartlett, 1950, 1951), Kaiser-Guttman rule or eigenvalues greater than one (EV1) (Guttman, 1954), parallel analysis (PA) (Horn, 1965), the scree test (Cattell, 1966), minimal average partial (MAP) (Velicer, 1976), and later modified parallel analysis (Glorfeld, 1995). These methods originated in the multivariate analysis area as tools to accompany either principal component analysis or exploratory common factor analysis. At the time of their development, these methods assumed a researcher predetermined an experimental model and after building an instrument based on that model, extracted a set of representative factors from the data. As a result of their original focus, the aforementioned methods for determining *k* tend to extract all of those factors that are above the "elbow" or change point in a scree plot. An implicit assumption of these methods is that the researcher is able to separate meaningful patterns from random noise (Jackson, 1993).

Within the text mining community, many of the aforementioned methods for determining *k* are still applied. Research into this subject continues mostly because these methods are manual in nature. Efron (2005) evaluated a number of the aforementioned methods and introduced a

variant of PA (Horn, 1965) called amended parallel analysis (APA). Likewise, Zhu and Ghodsi (2006) introduced a new method for selecting *k*, based on the profile likelihood. These methods were developed specifically for the text mining environment, however, like their predecessors, they attempt to find essentially the elbow point in a scree plot. In information retrieval, the issue of the optimal number of *k* is frequently detoured. Retaining an almost arbitrarily large number of factors is frequently the norm. Typically, retaining 300 factors or more is common particularly when the number of documents is high ($d \approx 10{,}000$) (Bradford, 2008). This or bigger values of *k* can be a reasonable if the target corpus is the entirety of the World Wide Web. However, in a corpus of 1000 or fewer documents, retaining only 10 factors is common (Bradford, 2008).

The R statistical package LSA does provide some dimensionality calculation routines (Wild, 2012). Of interest, it provides Kaiser dimensioning and an option to extract all observed (raw) dimensions. Additional techniques are available in the LSA package; however, they are not well supported in the literature. For example, dimensioning equal to the number of documents or a fraction of the singular values observed in the data.

Like LSA, LDA inherits this unresolved question of how many topics to extract. At the introduction of LDA, the issue of how many topics to extract was treated in an arbitrary manner extracting 100 topics in one example and 50 topics in another example. In the R statistics package *topic models*, two different methods of determining the number of topics to extract are provided (Grun and Hornik, 2011). The first is a 10-fold cross-validation method which can be set to test at many different intervals. The configuration of the 10-fold cross-validation as described by Grun and Hornik (2011) consisted of testing, for example, at 10, 20, 30, 40, 50, 100, and 200 topics with LDA with α estimated, LDA with α fixed, a LDA-Gibbs sampler, and

CTM for a total of 28 different runs. Only the best results of each run are retained based on the log-likelihood $\log(p(w|z))$ and then evaluated graphically looking for the overall best model (highest log-likelihood).

The 10-fold cross-validation test is time consuming. An alternative is also provided by running the log-likelihood test on the results from the LDA algorithm obtained by testing at each level of *k* over some interval or every other level of *k* over the same interval, e.g. test each value of *k* from 2-50, or test every other value of *k* from 2-100. This process is much faster and more thorough than testing at every 10^{th} *k*.

Likewise, NMF was introduced extracting an arbitrary 200 topics in an example (Lee and Seung, 1999). Brunet et al. (2004) introduced a cophenetic correlation coefficient, $p_k(\bar{C})$, with which the smallest values of *k* are selected before the magnitude of the coefficient begins to fall. A second method is the residual sum of squares (RSS) (Hutchins et al., 2008). RSS has the ability of not only estimating the appropriate number of topics to extract, it will also reveal if NMF analysis is appropriate to the corpus (Hutchins et al., 2008). In RSS, the effort is focused on variation in the RSS between the X matrix and WH matrices. A plot of the RSS shows an inflection when *r* strikes the proper number of dimensions (Hutchins et al., 2008). The R statistical package NMF provides tools for both of these dimensioning methods plus two additional tests *residuals* and *dispersion* (Gaujoux, 2010 and 2012). It is unclear if *residuals* refers to Frigyesi and Hoglund (2008) or Kim and Tidor (2003). Dispersion is not further identified.

While the optimal number of topics to extract with any one of the selected text mining methods represents a wide-open research opportunity, the immediate need of discovering the optimal number of topic to extract seems most precarious. Frequently in LSA the final solution

comes down to a judgment call with several solutions being generated each with different *k.* After examining the results, the researcher makes a subjective call as to which solution is best.

One alternative is to run each of the models and compare results for some common solution. This strategy, while interesting, has not been documented in the literature.

CHAPTER 3

TOPIC QUALITY EVALUATION

Introduction

The objective of the main experiment in this dissertation is to evaluate the perceived effectiveness of the text mining as posed by the research question; "which algorithms do humans perceive as more effective at extracting topics and classifying documents to topics?" Answering this question requires the development of a new theoretical model and a new data collection instrument. To assist in development of the new model, the following critical components of text mining are recapped: the process begins with a corpus or collection of documents. As described in chapter 2, low information words are removed from the corpus, the remaining words are cut back to their roots, and subsequently referred to as *terms*. A matrix is then extracted from the corpus that is composed of terms-by-documents. For each document of the corpus, a new column is added to the matrix, and the frequency of each term found in the document is recorded on a unique row of the column. Once a matrix is extracted, the text mining algorithm is applied to reveal the latent (hidden) topic structure. For any given latent topic, only certain terms define the topic and only select documents from the corpus are associated with it. Figure 5 presents a conceptual map of the text mining process.

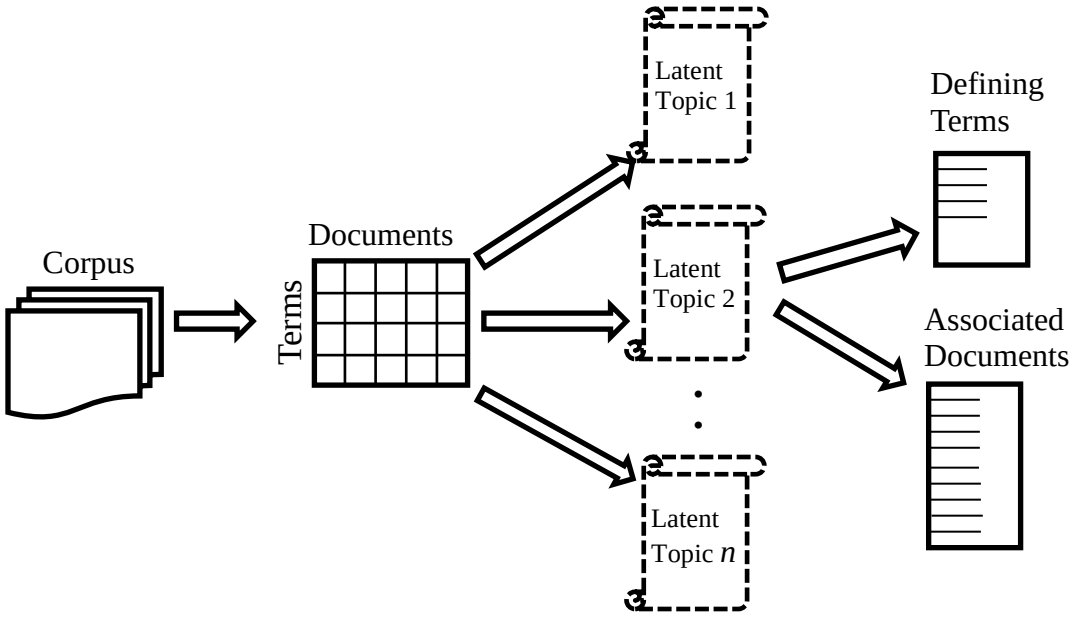

Figure 5. Illustration of the text mining process.

Defining the Theoretical Model Constructs

The preponderance of historical text mining algorithm evaluations has been quantitative in nature (see Table 1). An alternative approach is to use a qualitative experiment consistent with software engineering evaluation methods (Kitchenham, 1996). In such an experiment, groups of potential users function as software managers, developers, or maintainers and assess feature(s) of a software method or tool according to some predefined criteria (Kitchenham, 1996). These are formal experiments that require the potential user to make a subjective assessment along some evaluation criteria (Kitchenham, 1996).

In order for the potential user to conduct the subjective evaluation, some evaluation criterion is required. One possible criterion is how well the potential user understands, or how clear the subject matter of the latent topic is, based on the list of defining terms and list of associated documents. A second criterion is how clearly the associated documents reflect the latent topic. The third evaluation criterion then is to consider if the defining terms consistently

reflect the latent topic. A graphical representation of these criteria and their relationship to the text mining process is presented in Figure 6. A model consisting of such evaluation criteria was not found in the literature. However, while distinct, such criteria are similar to that used in document clustering algorithm testing (Kummamuru et al, 2004).

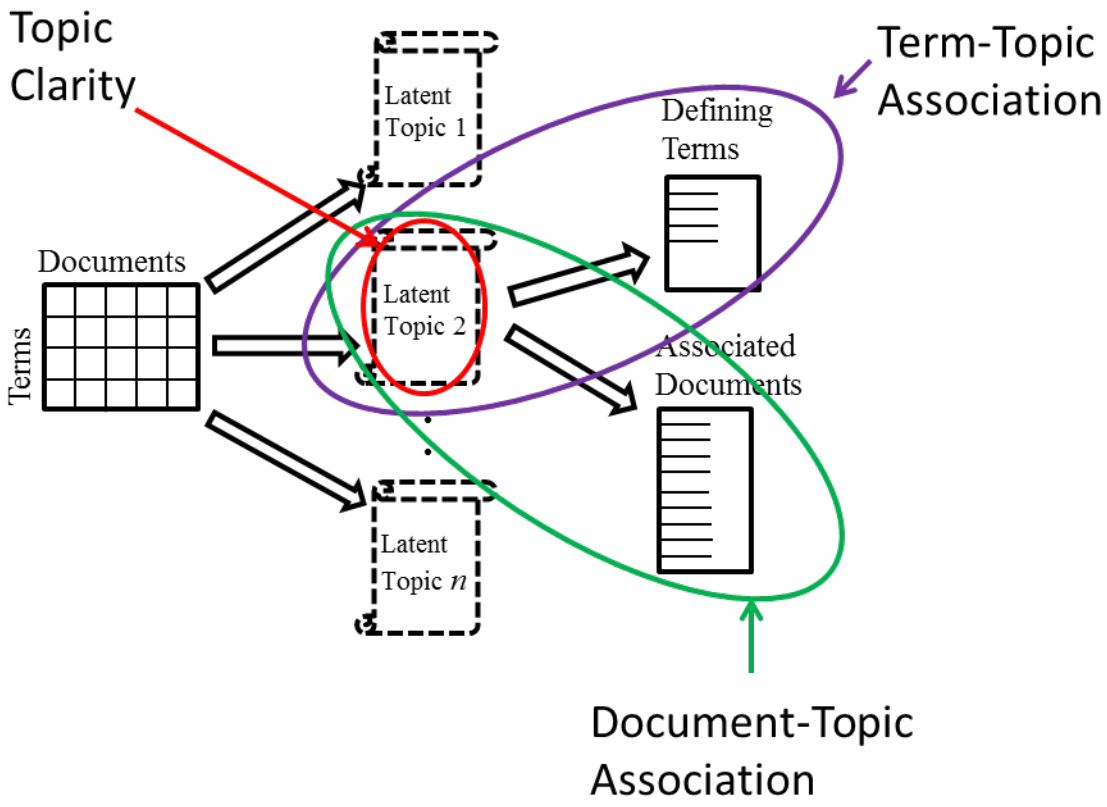

Figure 6. Evaluation criterion and text mining relationships.

Figure 7 presents the aforementioned measurement criterion as constructs in a theoretical research model. In this model, consistency with which the defining terms reflect the latent topic is retitled *Term-Topic Association*. The clarity with which the documents express the latent topic is titled *Document-Topic Association*. How clear the subject matter of the latent topic is titled *Topic Clarity*. The data collection instrument will also address the internal consistency of the documents through a fourth stand-alone construct called *Document Cohesion*. Internal

consistency of the terms is not addressed because term list entries are more representative of parts of speech, for example, a term list can consist of a mixture of nouns, verbs, adverbs, and adjectives. The order of assembly of these terms into a speech element may or may not be obvious to the casual observer and typically takes an experienced analyst some study time to interpret. In addition, in homogeneous corpora, writers routinely compose documents that discuss multiple topics. Therefore, a term list might include terms that describe the primary topic and a few terms that describe other topics in the corpus resulting in a list of terms that at times can be incoherent in isolation.

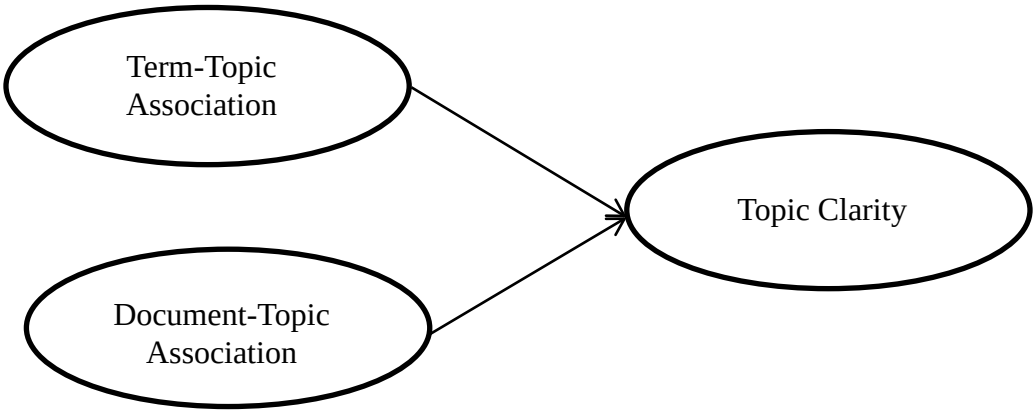

Figure 7. Theoretical research model.

CHAPTER 4

METHODS

Introduction

To address the research question introduced in Chapter 1, a 3 × 5 (analytic methods × discussion topics) factorial design experiment was performed. The data set used for this experiment originates from a real world business and possesses an unknown structure. The corpus was analyzed by LDA, LSA, and NMF with the results subsequently evaluated by potential user subjects.

Description of Experimental Data

Data used in the experiment consists of homogeneous group of documents focused on customer satisfaction; however, the sub-topic structure of the data is unknown. This data comes from a Fortune 500 retailer that offers an online service to customers. This corpus was generated by customers of the firm that had decided to cancel their subscription services with the provider during a single day in April 2006. As a step in the cancellation process, customers were asked why they decided to cancel their subscription and were given an open-ended text box in which to type their response. During data collection, the system was configured to force the customer to place something in the text box. This data is not a reduced sample and is nearly a census of activity for that day. No other reference cues were provided.

Since the firm is a member of a highly competitive industry, alterations were made to the analytic results to prevent revealing the provider's identity or its competitive industry. Descriptive words replace actual service and product names, e.g. [Product], to protect the firm's

identity and its specific industry segment. Select figures that could disclose the identity of the firm are modified by proportioning allowing a meaningful ratio to remain.

This data set consists of 1,143 customer comments. Many of the documents address service quality issues as a reason for departure from the firm; however, there are also a number of other issues that customers express, that exceed the service quality domain or are otherwise beyond the firm's ability to control. This corpus is a mixture of homogeneous and heterogeneous topics and is representative of a typical business corpus.

Analytic Software and Extraction of the X Matrix

Throughout the text mining algorithm literature, when competitive testing is conducted, methods are compared using the same corpus. However, an inconsistency in the literature is the programming platforms on which the algorithms are written. Frequently, one algorithm will be written under one programming language, e.g. Python, MatLab, Java, C, or C++ while the competing method is written in another language. While this is not necessarily detrimental or advantageous to the main analytic algorithm, text mining involves many preprocessing or data preparation steps. The literature is not clear as to the equality of the X matrices that the algorithms are analyzing. In this research, all analysis was conducted using the R statistical programming language (R Development Core Team, 2008). All preprocessing was conducted using the Text Mining (tm) package, which is a library of software tools for text mining (Feinerer, Hornik, and Meyer, 2008; Feinerer 2008, 2011, and 2012). These preprocessing steps were described in the Common Processing section of Chapter 2. Once preprocessing for each of the three data sets was completed, algorithm based analysis was performed against the X matrix by the LSA package (Wild, 2007 and 2012), NMF package (Gaujoux and Seoighe, 2010;

Gaujoux 2010 and 2012), and the Topic Models package (Grün and Hornik, 2010, 2011, and 2012). The results generated in that analysis were then applied to the main experiment.

Experimental Setup

The research question is addressed with an experimental design utilizing the analysis results obtained from the algorithms. The experimental design utilizes a randomized block factorial design (RBF-35) that considers two treatments consisting of the text mining method and the topics extracted. There are 15 total treatment combinations composed out of three text-mining methods and five topics. Fifteen treatments are too much to present to a single respondent. Therefore, each respondent evaluated 9 nine randomly selected treatments out of the 15 possible.

The number of topics to consider in this experiment was subjectively set at five. Considering that corpora routinely consist of tens to hundreds of subtopics, evaluating the methods on fewer topics seems unreasonable. However, the evaluation process as conceived is time consuming on the part of the respondent and the desire was to balance the information gain versus respondent payoff. Additionally, the corpus, when analyzed by singular values consists of six topics prior to the first change point (Zhu and Ghodsi 2006). Of those six topics, five appear in all three solutions. The sixth topic was discovered by LSA and NMF; however, LDA's solution had a unique sixth topic. Therefore, this sixth topic was dropped because it would throw the analysis out of balance.

While the three analytic methods were able to extract the same five discussion topics, those topics were not extracted in the same order. During a review of the results, the topics were renumbered so that topic number *n* was the same for all three methods. The order in which the

topics are numbered relates to the order in which matching topics were discovered across the methods and does not reflect the order of extraction.

For all algorithms, some topics will extract very cleanly while others possess high levels of noise (misclassified documents). When topics are extracted, the first topic always explains the highest level of variance or in the case of LDA the first topic has the highest probability. Then the level of variance explained decreases with each subsequent topic. Further, within a topic, the first document listed in the table of associated documents has the highest level of correlation to the topic. As one moves down the list of associated documents, the correlation level decreases. How much this decrease affects the clarity of the topic or clarity of the association's is unclear and undocumented in the literature. As a method of looking at this issue, the documents that were presented to respondents were either the ten highest correlated documents in the topic (odd numbered topics) or were 10 documents selected beginning at the 50^{th} percentile (even numbered topics).

Instrument Development

The initial instrument was developed and tested in two pilot studies during the summer of 2012. Following those pilots, the instrument was revised to its final form consistent with the theoretical research model of Figure 7. An instrument template of the final model is available in Appendix B. The entire model was not provided in Appendix B because the main research questions are designed into the full instrument 15 times, once per topic. See the section *Sample of Presented Data* in Appendix B for an example of 1 of the 15 topics that composed the full instrument.

The term-topic construct measures how well the respondents perceive the terms as defining the discussion topic. In this construct, references to "terms" are replaced with "keywords". Since the potential users were not trained in text mining, the objective of this word replacement was to promote clarity and understanding in the research instrument. The instruments four constructs are listed below followed by their items:

Term-Topic Association Construct

A1. The keywords accurately reflect the topic being discussed.

A2. After examining the keywords, it was easy to understand what the topic was about.

A3. The keywords are helpful in understanding the topic.

A4. The keywords define a single topic.

A5. The keywords are related to each other.

Document-Topic Association Construct

B1. The documents accurately reflect the topic being discussed.

B2. After examining the documents, it was easy to understand what the topic was about.

B3. The documents are helpful in understanding the topic.

B4. The documents define a single topic.

B5. The documents are related to each other.

Topic Clarity Construct

C1. The discussion topic is clear to me.

C2. The concept of the topic is clear.

C3. I feel I know what this topic is about.

C4. I would be able to label this topic.

{{Provide a text box here and instruct the participant to label (define) the topic.}}

C5. How easy was it to label this topic?

Document Cohesion Construct

D1. All these documents talk about the same thing.

D2. The documents address a topic in a consistent manner.

D3. How easy is it to identify documents that "do not belong" in the group?

D4. These documents are similar to each other.

Sources of Variation

In analyzing the solutions, variation was anticipated from the following sources:

- Text mining method
- Topic – because some corpora possess topics with highly focused discussion subjects while other topics are broader in scope.
- Presentation order – participants are required to make an evaluation which entails a degree of learning effect.
- Corpus – the scope of some corpora are highly defined while others are loosely defined. While not fully understood, topic homogeneity may also cause variation.
- Participants

These sources of variation are addressed in the following manners: corpus variations is controlled by *fixing;* the presented data are from one corpus; to control presentation order

variation, the nine treatment combinations presented to participants were in a *completely randomized* order*;* finally, text mining method, topic, and participant are *accounted for* in the ANOVA model.

Threats to Validity and Generalizability

Four threats to validity are identified in the design and are addressed as defined here. Linguistic ability of participants is an important threat because participants are asked to make judgments of linguistic results. Lower levels of English language ability will impede the respondent's ability to provide accurate judgments of accuracy or cohesiveness. A demographic question is added to the instrument asking about the respondent's first language. Significance of the item is tested with the ANOVA procedure.

Previous exposure to text mining methods could result in a bias of the results. A respondent with professional experience in data analysis may have a preexisting preference for one of the particular method and consciously or subconsciously skew their ratings. At no point does the instrument provide the participant with method information. Further, the results that are presented in the instrument do not emulate the output of the software package or analytic method that a user could have potentially experienced. Finally, a demographics question asks about the participant's familiarity with text mining tools methods, and algorithms.

In some research, student subjects can be problematic for generalizability. For instance, they have been found to provide more homogeneous responses in marketing research (Peterson, 2001). In other studies, the results are not as conclusive yet suggest student responses are generalizable (Enis et al., 1972). In a real world business environment, text-mining operators require a degree of expertise; however, in this experiment the student is not asked to manipulate

the data, instead, they are simply asked to judge the understandability of the results. Those judgments require an ability to interpret language material. To further measure the student's ability with this research, the demographics section the instrument also collects information on first language, the number of semester hours completed in language, and how well they understand the documents.

Since respondents are exposed to all of the treatment combinations, a learning effect is expected. To detect it and eliminate it if necessary, treatments are presented in random order and the presentation order for each participant is captured.

CHAPTER 5

DATA ANALYSIS

Introduction

Analysis of the survey instrument data follows the traditional multivariate procedures. Analytic procedures included factor analysis; development of and evaluation of summated construct items; evaluation of order effect; evaluation for an effect on the summated constructs caused by the analytic method and topic; finally, regression analysis of the main experimental model shown in Figure 7 as well as regression analysis with demographic data.

Data Collection and Cleanup

The concept behind the experiment is to evaluate the perceived effectiveness of text mining algorithms as posed by the research question:

Which algorithms do humans perceive as more effective at extracting topics and classifying documents to topics?

To answer this question, a group of potential users were presented analysis samples from each of three text mining methods and asked to evaluate them using a Likert scale questionnaire. The measurement instrument used for this experiment is contained in Appendix C. This experiment utilized data set 3, customer comments data, which is described in chapter 3. Each of the three text mining methods extracted six discussion topics from the data which was in turn loaded into the research questionnaire. The questionnaire was administered by Qualtrics online survey software (www.qualtrics.com). To avoid respondent fatigue, Qualtrics was configured to randomly select 9 of the 18 possible solutions to present to each respondent. Randomization coding was retained for analysis. Three hundred fifty four respondents participated in the study.

During the survey collection period, data from 354 respondents was collected. Five of the respondents were deleted because they terminated shortly after the informed consent and no usable data was collected from them. Of the 5 respondents, 1 declined to participate beyond the informed consent, 2 terminated immediately after the informed consent, and 2 terminated their participation during the introduction. This left 349 respondents each of which experienced 9 combinations for a total of 3,141 observations (349 × 9). This total of 3,141 observations was possible because each respondent evaluated nine sets of text mining results. In several instances, respondents skipped a set of questions for one construct. Since the analysis did not use paired samples, those empty observations were deleted. Once the incomplete or blank observations were removed, 2,542 observations remained for analysis. During the data review, the demographics variable "competence at categorizing text" was reverse coded. This variable had been set up with option 1 = *very competent* through 4 = *very challenged*. All other variables were configured low-to-high. One additional variable was added to the data set, *order*. The Qualtrics output included a segment of data with each respondent that identified the order that the treatments were presented based on the block title found in the Qualtrics instrument. Order is a simple numerical value based on the Qualtrics randomization data that identifies what order (1 through 9) the observation was presented to the respondent. All measurement item titles were recoded with an alpha-numeric code that identified the construct of association and the item number within that construct. Table 2 provides the item text and codes used in subsequent analysis for each construct.

In another odd event, Qualtrics actually presented 12 sets of text mining results to respondent 210. All 12 observations were retained; however, for order effect testing, the three extra observations were deleted because their presentation order was not recorded by Qualtrics.

Table 2

Construct Items and Coding used in Subsequent Analysis

Code	Item Text
Term-Topic Association Construct	
A1	The keywords accurately reflect the topic being discussed.
A2	After examining the keywords, it was easy to understand what the topic was about.
A3	The keywords are helpful in understanding the topic.
A4	The keywords define a single topic.
A5	The keywords are related to each other.
Document-Topic Association Construct	
B1	The documents accurately reflect the topic being discussed.
B2	After examining the documents, it was easy to understand what the topic was about.
B3	The documents are helpful in understanding the topic.
B4	The documents define a single topic.
B5	The documents are related to each other.
Topic Clarity Construct	
C1	The discussion topic is clear to me.
C2	The concept of the topic is clear.
C3	I feel I know what this topic is about.
C4	I would be able to label this topic.
C5	How easy was it to label this topic?
Document Cohesion Construct	
D1	All these documents talk about the same thing.
D2	The documents address a topic in a consistent manner.
D3	How easy is it to identify documents that “do not belong” in the group?
D4	These documents are similar to each other.

During the initial attempt at testing for an analytic effect of *Method* and *Topic* on the summated constructs, it was discovered the data contained an error. After extensive review, it was discovered that Topics 2 and 7 should have been deleted. Topic 2 was only discovered, as defined by its term structure, by the analytic methods LSA and NMF. This topic, as defined by its terms was not observed in the LDA results. However, LDA found another topic, which was referred to as Topic 7. With these two topics included in the analysis, the respective cells are not filled across all methods and therefore cause problems in the analytic results. These topics were

removed from the analysis data. This resulted in 276 observations against Topic 2 and 147 observations against Topic 7 being deleted.

After deleting observations for Topics 2 and 7, there were 2,118 observations in the data set. Fourteen observations were deleted because they were missing data for the *Topic Clarity*, *Term-Topic* or *Document-Topic* construct. With these observations deleted, there were 2,104 observations available for analysis.

After this first round of cleaning, the data was moved into SPSS Statistics 20 for further cleaning based on residual analysis. This action is taken to remove outlier and influential observations from the dataset. With outliers and influential observations removed, analysis proceeded in SPSS and consisted of seven sequential steps consistent with multivariate methodology (Hair et al., 2006, ch. 3 and 4). Those steps were factor analysis of items into constructs; internal consistency testing of constructs; testing of the order effect on each of the constructs by Univariate analysis; evaluation of the predictive ability of text mining method and topic on the constructs; multiple regression of the main model; integration of demographic information into the main model; and finally, residual analysis. Each of these steps and the discoveries made are presented in the sections that follow.

Residual Analysis

Residuals were examined by standardized residuals, leverage, and Mahalanobis D^2 statistic while considering the regression of the dependent variable *Topic Clarity* onto the two independent variables *Document-Topic Association* and *Term-Topic Association*. Outlier detection first considered the standardized residuals generated by the linear regression process. Because the sample is large (greater 2,000), a guideline of ± 3 standard

deviations was adopted as a guide with any observation greater considered an outlier and therefore deleted (Hair et al., 2006, p. 75). Residuals were next evaluated with Mahalanobis' statistic using the *p*-value as generated by using SPSS' χ^2 statistic and the equation p = 1-cdf.chisq(x_i,2). The decision rule was to consider an observation with a p-value $\leq .001$ as an influential observation and were deleted. The linear regression and residual analysis was then repeated until all observations exhibited standardized residuals < 3 and a *p*-value for the Mahalanobis statistic > .001 (Hair, 2006, p. 75). This resulted in deleting 109 observations or 5.1% of all observations. Therefore, deleted observations was not excessive (Burke, 2001). Final residual diagnostic statistics for the 1,995 observations are included in Table 3.

Table 3

Regression Model Residual Analysis

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1.14	6.83	4.97	1.172	1995
Std. Predicted Value	-3.269	1.590	.000	1.000	1995
Standard Error of Predicted Value	.016	.053	.026	.009	1995
Adjusted Predicted Value	1.14	6.84	4.97	1.172	1995
Residual	-1.987	2.064	.000	.697	1995
Std. Residual	-2.849	2.958	.000	.999	1995
Stud. Residual	-2.850	2.959	.000	1.000	1995
Deleted Residual	-1.989	2.065	.000	.698	1995
Stud. Deleted Residual	-2.855	2.964	.000	1.001	1995
Mahal. Distance	.012	10.695	1.999	2.217	1995
Cook's Distance	.000	.011	.001	.001	1995
Centered Leverage Value	.000	.005	.001	.001	1995

a. Dependent Variable: TopicClarity

Statistical testing of residual normality and equal variance are not effective due to the sample size. Normality tests including Kolmogorov-Smirnov and Shapiro-Wilks are known to be sensitive when the sample size exceeding 1,000 observations (Hair et al., 2006, p. 82). Likewise, Levene's test used to test equality of variance in the residuals,

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} N_i (Z_i j - Z_{i.})^2}, \quad (19)$$

includes an *N* term in the numerator that makes it sensitive to large samples as well. Therefore, a histogram of the residuals is provided in Figure 8, and a scatter plot of standardized residuals is provided in Figure 9.

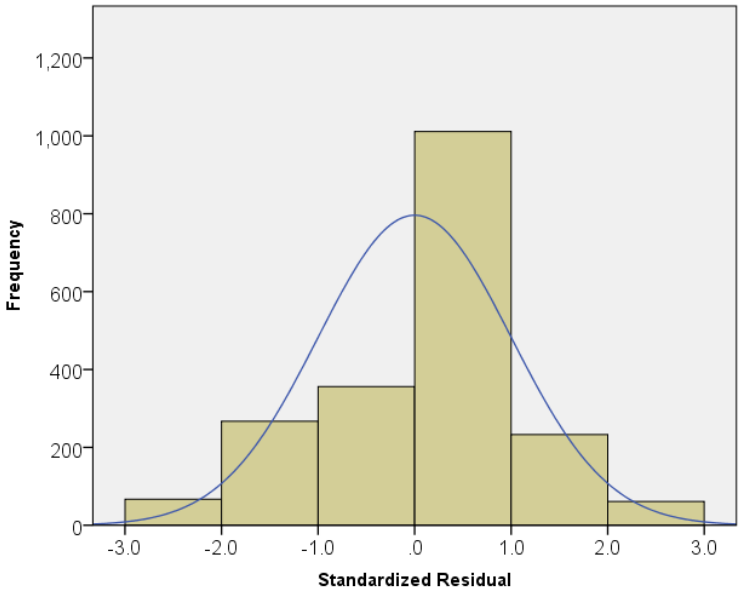

Figure 8. Histogram of topic clarity versus standardized residuals.

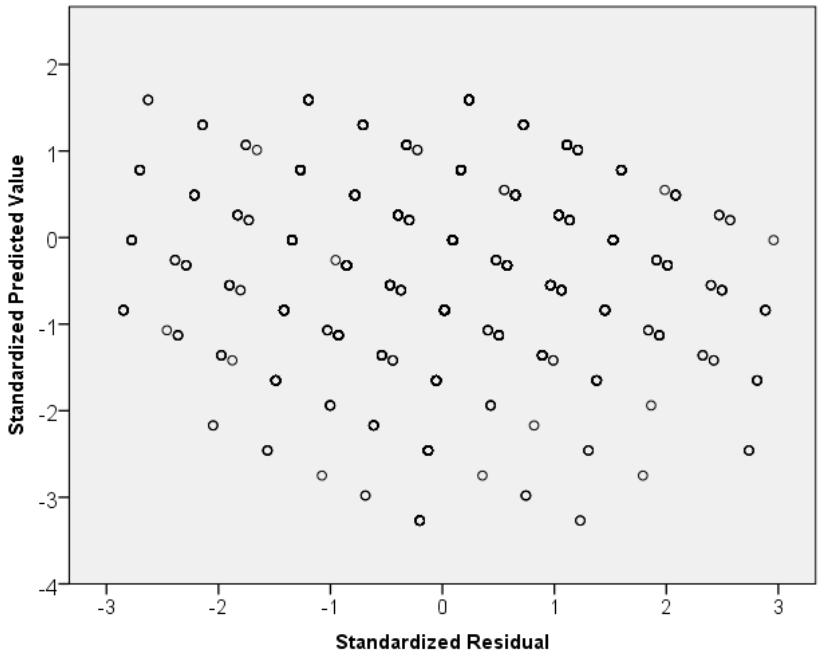

Figure 9. Scatterplot of predicted verses standardized residuals

Descriptive Statistics

Demographic characteristics associated with the 1,995 observations are contained in Figure 10 and Tables 4 through 7. The numbers listed in these Figures and Tables will not add up to 1,995 because a number of individuals did not provide various elements of the demographic data. Since demographic data is not critical to the main analytic effort, these observations were retained. Demographic characteristics collected included in Figure 10 are gender (Panel A and Table 4), age (Panel B and Table 5), education level (Panel C and Table 6), and whether English was the observers First Language (Panel D and Table 7).

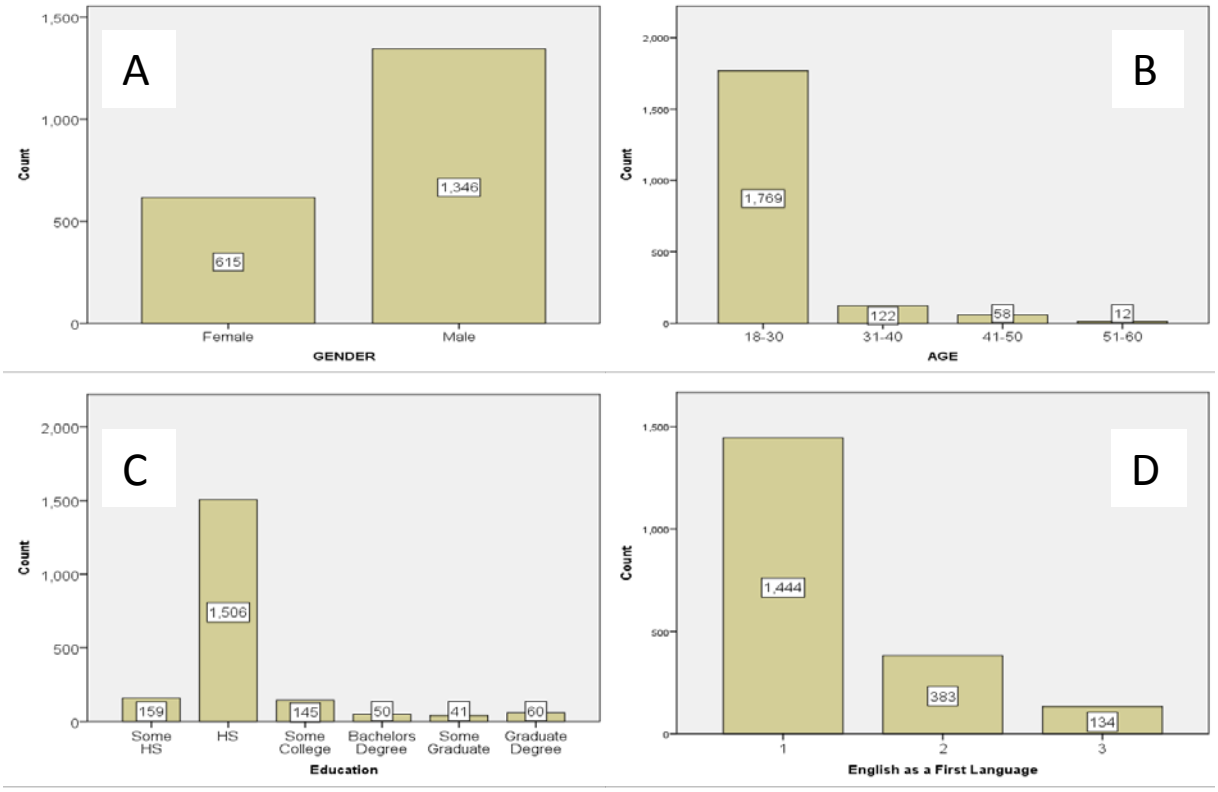

Figure 10. Survey participant demographic data

Table 4

Gender Demographic Frequency Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	615	30.8	31.4	31.4
	Male	1346	67.5	68.6	100.0
	Total	1961	98.3	100.0	
Missing		34	1.7		
Total		1995	100.0		

Table 5

Age Demographic Frequency Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18-30	1769	88.7	90.2	90.2
	31-40	122	6.1	6.2	96.4
	41-50	58	2.9	3.0	99.4
	51-60	12	.6	.6	100.0
	➢ 61	0	0	0	
	Total	1961	98.3	100.0	
Missing		34	1.7		
Total		1995	100.0		

Table 6

Education Demographics Frequency Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Some H.S.	159	8.0	8.1	8.1
	H.S. Graduate	1506	75.5	76.8	84.9
	Some College	145	7.3	7.4	92.3
	Bachelors	50	2.5	2.5	94.8
	Some Graduate	41	2.1	2.1	96.9
	Graduate	60	3.0	3.1	100.0
	Total	1961	98.3	100.0	
Missing		34	1.7		
Total		1995	100.0		

Table 7

English as a First Language Frequency Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	1444	72.4	73.6	73.6
	No.Fluent	383	19.2	19.5	93.2
	No Learning	134	6.7	6.8	100.0
	Total	1961	98.3	100.0	
Missing		34	1.7		
Total		1995	100.0		

This experiment depends on the respondent possessing a degree of language comprehension skills. To assess the respondent's skill and task understanding, additional self-reported task specific information was collected. These items are reported in Figure 11 and include how familiar the respondent was with text mining prior to this experiment (Panel A and Table 8). The number of semester hour of completed language courses (Panel B and Table 9). How well the respondent felt they understood the documents presented to them was also asked and the results are reported in Panel C and Table 10. Finally, how competent the respondent felt they were at categorizing text into classes was also asked. Those results are reported in Panel D and Table 11.

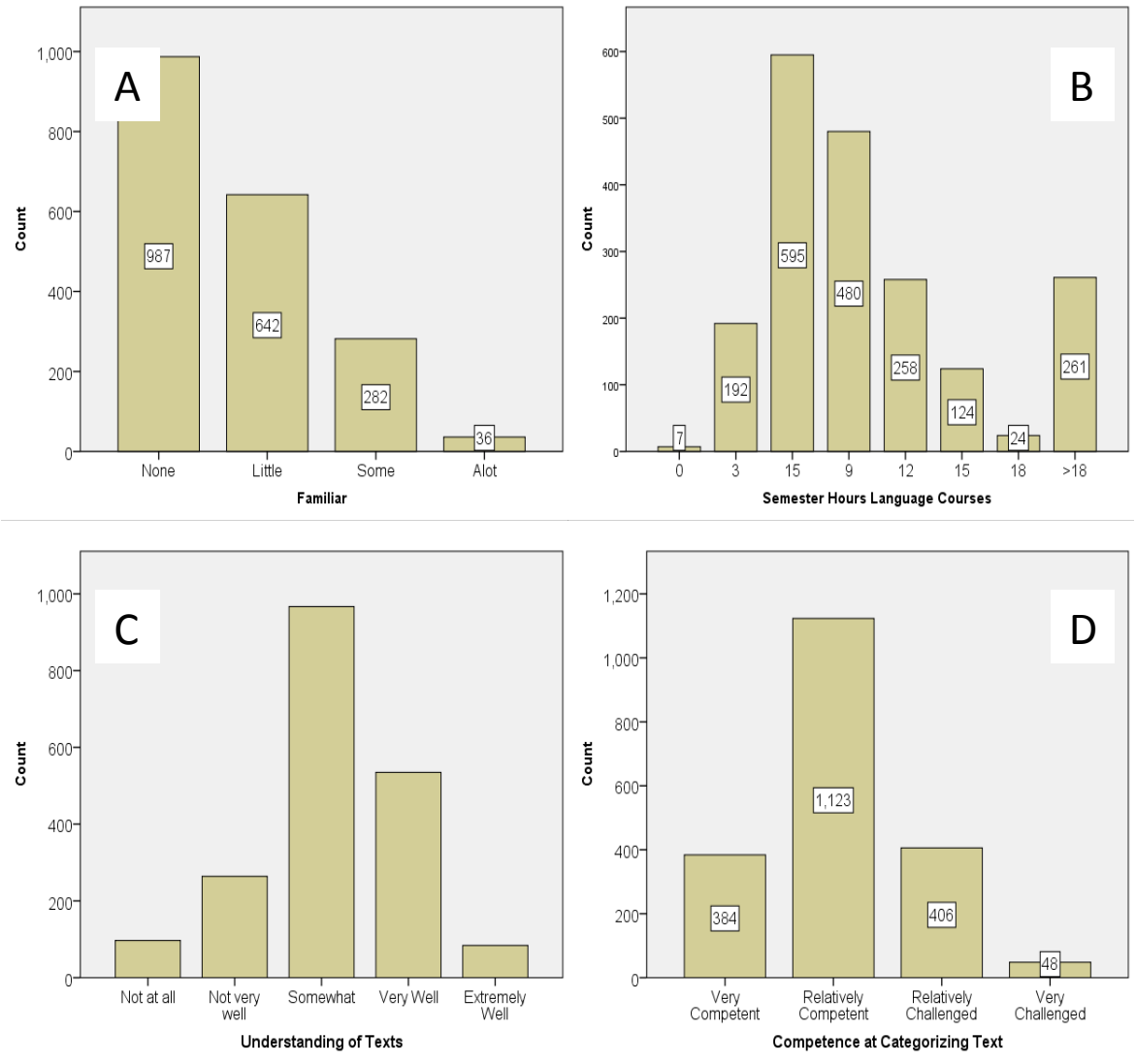

Figure 11. Self-reported task competence measures.

Table 8

Familiarity with Text Mining

Familiarity		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	None	987	49.5	50.7	50.7
	Little	642	32.2	33.0	83.7
	Some	282	14.1	14.5	98.2
	Alot	36	1.8	1.8	100.0
	Total	1947	97.6	100.0	
Missing		48	2.4		
Total		1995	100.0		

Table 9

Semester Hours of Language Courses Completed

Hours		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	7	.4	.4	.4
	3	192	9.6	9.9	10.3
	6	595	29.8	30.7	40.9
	9	480	24.1	24.7	65.6
	12	258	12.9	13.3	78.9
	15	124	6.2	6.4	85.3
	18	24	1.2	1.2	86.6
	➢ 18	261	13.1	13.4	100.0
	Total	1941	97.3	100.0	
Missing		54	2.7		
Total		1995	100.0		

Table 10

Understanding of the Presented Documents

Understanding		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not at all	97	4.9	5.0	5.0
	Not very well	264	13.2	13.6	18.5
	Somewhat	967	48.5	49.7	68.2
	Very well	535	26.8	27.5	95.7
	Extremely well	84	4.2	4.3	100.0
	Total	1947	97.6	100.0	
Missing		48	2.4		
Total		1995	100.0		

Table 11

Competence at Categorizing Text into Classes

Competence		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very competent	384	19.2	19.6	19.6
	Relatively competent	1123	56.3	57.3	76.8
	Relatively challenged	406	20.4	20.7	97.6
	Very challenged	48	2.4	2.4	100.0
	Total	1961	98.3	100.0	
Missing		34	1.7		
Total		1995	100.0		

Analysis

Step 1: Factor Analysis

The objective of the first step of the analysis was to verify that the structure of the items, and reduce the items into a smaller set of composite variates or factors (Hair et al., 2006, p. 107). Factor analysis was performed by principal components analysis extracting four a priori factors (Hair et al., 2006, p. 120). Initially, VARIMAX rotation was used since it is a widely popular rotation method. However, while these results do exhibit the a priori constructs, the results showed a number of significant cross loadings with factor loading above 0.40 as shown in Table 12.

Table 12

Principal Component Analysis with VARIMAX Rotation Utilizing a .40 Suppression Level

	Component			
	1	2	3	4
A1	.710			
A2	.736			
A3	.732			
A4	.725	.447		
A5	.666	.454		
B1			.403	.658
B2			.460	.646
B3			.407	.687
B4		.537		.563
B5		.496		.578
C1			.660	.405
C2			.682	.410
C3			.709	
C4			.703	
C5		.419	.710	
D1		.765		
D2		.714	.404	
D3		.655	.403	
D4		.762		

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 9 iterations.

QUARTIMAX and EQUAMAX rotations showed similar if not worse results. Brown (2009a) quotes Tabachnick and Fiddell (2007, p. 646):

> *Perhaps the best way to decide between orthogonal and oblique rotation is to request oblique rotation [e.g., direct oblimin or promax from SPSS] with the desired number of factors [see Brown, 2009b] and look at the correlations among factors…if factor correlations are not driven by the data, the solution remains nearly orthogonal. Look at the factor correlation matrix for correlations around .32 and above. If correlations exceed .32, then there is 10% (or more) overlap in variance among factors, enough variance to warrant oblique rotation unless there are compelling reasons for orthogonal rotation.*

VARIMAX rotation is one of the orthogonal rotation methods. Hair et al. (2006, p. 127) likewise argues that correlated factors are best analyzed by one of the oblique rotation methods. Following Tabachnick and Fiddell (2007, p. 646), the correlation matrix generated by OBLIMIN rotation presented in Table 13 shows all components exceeding the 0.32 level of correlation. This is not surprising given the special relationship of the constructs to one another. In natural language, word selection defines a latent discussion topic and documents that discuss the topic are composed of the words that defined it.

Table 13

OBLIMIN Component Correlation Matrix

Component	1	2	3	4
1	1.000	.347	.722	-.770
2	.347	1.000	.409	-.374
3	.722	.409	1.000	-.818
4	-.770	-.374	-.818	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Correlation was also tested with the PROMAX component correlation matrix. The results, presented in Table 14, are even stronger than those found using OBLIMIN rotation. The minimal correlation value found by PROMAX correlation is 0.690 while with OBLIMIN correlation the minimal correlation was 0.347. Additionally, with the OBLIMIN correlation matrix, one factor component is negatively correlated. An explanation of that negativity is not apparent from the basic research model.

Table 14

PROMAX Component Correlation Matrix

Component	1	2	3	4
1	1.000	.694	.720	.780
2	.694	1.000	.690	.698
3	.720	.690	1.000	.765
4	.780	.698	.765	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

Given the strength of the PROMAX correlation results, principal components factor analysis was ran again with PROMAX rotation. The final factor results are excellent and presented in Table 15. This solution exhibits no cross-loading items and the minimal factor loading value is 0.612. Only one loading is below the 0.7 level, which is considered a well-defined structure (Hair et al., 2006, p. 128). Attempts at factor analysis using OBLIMIN rotation resulted in very weak loadings in the "D" (Document cohesion) construct and cross-loading onto "C" (topic clarity). When these items were removed, the result became worse with more items then exhibiting cross-loading.

Table 15

Principal Components Factor Analysis Results using PROMAX Rotations

	Component			
	1	2	3	4
TermTopicAssociation4	.849			
TermTopicAssociation2	.815			
TermTopicAssociation3	.806			
TermTopicAssociation1	.763			
TermTopicAssociation5	.716			
DocumentCohesion1		.831		
DocumentCohesion4		.811		
DocumentCohesion2		.718		
DocumentCohesion3		.647		
TopicClarity5			.840	
TopicClarity3			.749	
TopicClarity4			.748	
TopicClarity2			.682	
TopicClarity1			.639	
DocumentTopicAssoc3				.823
DocumentTopicAssoc1				.759
DocumentTopicAssoc2				.723
DocumentTopicAssoc4				.625
DocumentTopicAssoc5				.620

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 9 iterations.

Communalities, which measure the amount of variance accounted for by the factor solution are also provided. A generally accepted guideline for communality measurement is that that at least half (0.50) of the variable's variance is accounted for. Communalities are reported in Table 16. The minimal observed communality it 0.745.

Table 16

Communalities

	Initial	Extraction
A1	1.000	.829
A2	1.000	.868
A3	1.000	.849
A4	1.000	.815
A5	1.000	.788
B1	1.000	.846
B2	1.000	.863
B3	1.000	.858
B4	1.000	.793
B5	1.000	.815
C1	1.000	.868
C2	1.000	.877
C3	1.000	.874
C4	1.000	.859
C5	1.000	.775
D1	1.000	.837
D2	1.000	.848
D3	1.000	.745
D4	1.000	.866

Extraction Method: Principal Component Analysis.

Step 2: Summated Scales

With a satisfactory factor structure established, summated scale items were generated by average scores. The new summated variables were named consistent with the constructs they represent. Reliability (internal consistency) of the new scales is measured by Cronbach's alpha. All constructs have excellent internal consistency with Cronbach's $\alpha > 0.90$ (see Table 17).

Table 17

Reliability Testing of Summated Scale Items

Construct Title	**Items**	**Cronbach's α**
Term-Topic Association	A1-A5	.942
Document-Topic Association	B1-B5	.943
Topic Clarity	C1-C5	.955
Document Cohesion	D1-D4	.935

Step 3: Order Effect

During summer 2012, preliminary pilot studies suggested an order effect might be present. During that study, two instruments were used that presented treatments in a fixed order. The source of the order effect was not clear, however, a learning effect was suggested. To reduce the impact of an order effect on the study results, if it were present, this study utilized a fully randomized presentation order. Here I want to test and make sure that presentation order is not significant for any of the four summated constructs. This was tested by using a Univariate analysis procedure. The test hypothesis for this procedure is specified below:

H_0: Order does not affect the dependent variable

H_A: Order does affect the dependent variable

Each of the four summated constructs were tested for order effect individually. None exhibited an order effect at the .05 significance level. Result are presented in Tables 18 - 21. All *p*-values are 0.4 and larger. The decision is to fail to reject the null and conclude the data does not show signs of an order effect on any of the constructs. Notice that the total degrees of freedom have been reduced by two because of the extra observations generated by respondent 210 who was presented 12 topics for evaluation. One of those three extra observations was deleted as an either an influential observation or an outlier and the remaining two were deleted for this analysis.

Table 18

Order Effect Testing for the Term-Topic Association Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3.560[a]	8	.445	.266	.977
Intercept	49544.316	1	49544.316	29597.166	.000
Presentation Order	3.560	8	.445	.266	.977
Error	3321.126	1984	1.674		
Total	52900.000	1993			
Corrected Total	3324.686	1992			

a. R Squared = .001 (Adjusted R Squared = -.003)

Table 19

Order Effect Testing for the Document-Term Association Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.868[a]	8	.984	.604	.775
Intercept	51043.088	1	51043.088	31348.739	.000
PresentOrder	7.868	8	.984	.604	.775
Error	3230.417	1984	1.628		
Total	54311.000	1993			
Corrected Total	3238.285	1992			

a. R Squared = .002 (Adjusted R Squared = -.002)

Table 20

Order Effect Testing for the Topic Clarity Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14.802[a]	8	1.850	.996	.437
Intercept	49213.852	1	49213.852	26486.250	.000
Presentation Order	14.802	8	1.850	.996	.437
Error	3686.452	1984	1.858		
Total	52938.000	1993			
Corrected Total	3701.253	1992			

a. R Squared = .004 (Adjusted R Squared = .000)

Table 21

Order Effect Testing for the Document Cohesion Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9.847[a]	8	1.231	.685	.705
Intercept	48027.001	1	48027.001	26731.517	.000
Presentation Order	9.847	8	1.231	.685	.705
Error	3560.947	1982	1.797		
Total	51621.000	1991			
Corrected Total	3570.794	1990			

a. R Squared = .003 (Adjusted R Squared = -.001)

Step 4: Testing for Analytic Method and Topic Effect

With presentation order demonstrated as insignificant, Univariate analysis of Analytic Method and Extracted Topic variables with the four experimental model constructs set as dependent variables was performed. The objective of this phase of the analysis was to determine if one analytic method performed better than another did. It also tested if any of the extracted topics helped explain the variation observed in the constructs. Once Univariate analysis was complete, the Topic and Method variables were tested by post hoc procedures to determine which discussion topic and which analytic method performed best. Finally, interaction between Topic and Method was explored by post hoc analysis.

Univariate Analysis

Since order is not a significant predictor, two observations for respondent 210 were reloaded to the data. Results for Univariate testing of the Method and Topic are reported in Tables 22 through 25. This analysis considered the analytic method and the topic and an interaction term as independent variable on to which the four main model constructs were regressed. The discussion topic and interaction term Method × Topic was for all four of the

constructs at the .05 significance level. The analysis method was found significant at the .05 significance level for the *Document-Topic Association* construct only. For the other constructs, the *p*-values run in the range of 0.053 and 0.079. While these constructs could be argued as marginally significant, they are not explanatory. When evaluated by the R^2_{adj} statistic, the analytic method only explains 2.2% of the variation in the Document Cohesiveness variable. Coefficient of determination and adjusted coefficient of determination for each construct are also reported in tables 22 through 25.

Table 22

Univariate Analysis of the Term-Topic Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	86.289[a]	14	6.164	3.764	.000
Intercept	49583.189	1	49583.189	30277.982	.000
Method	9.608	2	4.804	2.934	.053
Topic	44.912	4	11.228	6.856	.000
Method * Topic	31.964	8	3.996	2.440	.013
Error	3242.446	1980	1.638		
Total	52974.000	1995			
Corrected Total	3328.735	1994			

a. R Squared = .026 (Adjusted R Squared = .019)

Table 23

Univariate Analysis of the Document-Topic Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	94.215[a]	14	6.730	4.233	.000
Intercept	51061.157	1	51061.157	32117.738	.000
Method	10.772	2	5.386	3.388	.034
Topic	38.882	4	9.721	6.114	.000
Method * Topic	44.831	8	5.604	3.525	.000
Error	3147.827	1980	1.590		
Total	54385.000	1995			
Corrected Total	3242.042	1994			

a. R Squared = .029 (Adjusted R Squared = .022)

Table 24

Univariate Analysis of the Topic Clarity Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	98.566[a]	14	7.040	3.864	.000
Intercept	49217.626	1	49217.626	27011.557	.000
Method	10.237	2	5.119	2.809	.060
Topic	48.989	4	12.247	6.721	.000
Method * Topic	39.677	8	4.960	2.722	.006
Error	3607.748	1980	1.822		
Total	53003.000	1995			
Corrected Total	3706.314	1994			

a. R Squared = .027 (Adjusted R Squared = .020)

Table 25

Univariate Analysis of the Document Cohesiveness Construct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	147.859[a]	14	10.561	6.095	.000
Intercept	48041.549	1	48041.549	27726.280	.000
Method	8.787	2	4.394	2.536	.079
Topic	63.816	4	15.954	9.208	.000
Method * Topic	75.593	8	9.449	5.453	.000
Error	3427.297	1978	1.733		
Total	51695.000	1993			
Corrected Total	3575.156	1992			

a. R Squared = .041 (Adjusted R Squared = .035)

Analysis of the Topic Variable

To determine the mean differences among the topics, post hoc analysis with the Scheffé test was performed testing each of the extracted discussion topics against each summated construct. Scheffé was selected for this task because it is considered a more conservative statistic. Those results are reported in Tables 26 through 29. In each of the four constructs, at the .05 significance level, the 4th and 6th Topics are not significantly different from one another. Further, the 1st, 3rd, and 5th Topics are likewise not significantly different from one another.

However, these two subgroups are different with Topics 1, 3, and 5 possessing higher mean values than Topic 4 and 6. Additionally, within the odd numbered topics for most constructs, the means rank order is 3, 1, and 5. This might be coincidental although, the means rank order for the even numbered topics is in order with 4 possessing a higher mean than the 6th Topic. In general, I consider this coincidental because this numbering of topics is not the order of extraction. The result from each algorithm was compared one topic at a time against the other algorithms. As singular recurring topics were discovered across the three results they were numbered. Therefore the topic numbers observed here are arbitrarily assigned. As for the even numbered topics, it makes sense that they exhibit significantly lower means. Documents for those topics were selected from the fiftieth percentile of the association list while the odd numbered topics were selected beginning at the first percentile.

Table 26

Scheffé Testing of Topics on the Term-Topic Association Construct

(I) Topic	(J) Topic	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	3	.02	.091	1.000	-.26	.30
	4	.27	.091	.059	-.01	.55
	5	.12	.090	.755	-.15	.40
	6	.39*	.090	.001	.11	.67
3	1	-.02	.091	1.000	-.30	.26
	4	.25	.092	.111	-.03	.53
	5	.10	.091	.872	-.18	.38
	6	.37*	.091	.003	.09	.65
4	1	-.27	.091	.059	-.55	.01
	3	-.25	.092	.111	-.53	.03
	5	-.15	.091	.600	-.43	.13
	6	.12	.091	.802	-.16	.40
5	1	-.12	.090	.755	-.40	.15
	3	-.10	.091	.872	-.38	.18
	4	.15	.091	.600	-.13	.43
	6	.27	.090	.066	-.01	.54
6	1	-.39*	.090	.001	-.67	-.11
	3	-.37*	.091	.003	-.65	-.09
	4	-.12	.091	.802	-.40	.16
	5	-.27	.090	.066	-.54	.01

Based on observed means. The error term is Mean Square(Error) = 1.638.

Table 27

Scheffé Testing of Topics on the Document-Topic Construct

(I) Topic	(J) Topic	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	3	-.03	.089	.999	-.31	.25
	4	.27	.090	.063	-.01	.54
	5	.07	.088	.950	-.20	.35
	6	.31*	.089	.015	.04	.58
3	1	.03	.089	.999	-.25	.31
	4	.30*	.090	.029	.02	.58
	5	.10	.089	.850	-.17	.38
	6	.34*	.089	.006	.07	.62
4	1	-.27	.090	.063	-.54	.01
	3	-.30*	.090	.029	-.58	-.02
	5	-.19	.089	.325	-.47	.08
	6	.04	.090	.993	-.23	.32
5	1	-.07	.088	.950	-.35	.20
	3	-.10	.089	.850	-.38	.17
	4	.19	.089	.325	-.08	.47
	6	.24	.089	.129	-.04	.51
6	1	-.31*	.089	.015	-.58	-.04
	3	-.34*	.089	.006	-.62	-.07
	4	-.04	.090	.993	-.32	.23
	5	-.24	.089	.129	-.51	.04

Based on observed means. The error term is Mean Square(Error) = 1.590.

Table 28

Scheffé Testing of Topics on the Topic Clarity Construct

(I) Topic	(J) Topic	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	3	-.01	.096	1.000	-.30	.29
	4	.33*	.096	.018	.04	.63
	5	.12	.095	.826	-.18	.41
	6	.35*	.095	.009	.06	.64
3	1	.01	.096	1.000	-.29	.30
	4	.34*	.097	.014	.04	.64
	5	.13	.096	.785	-.17	.42
	6	.36*	.096	.007	.07	.66
4	1	-.33*	.096	.018	-.63	-.04
	3	-.34*	.097	.014	-.64	-.04
	5	-.22	.096	.279	-.51	.08
	6	.02	.096	1.000	-.28	.31
5	1	-.12	.095	.826	-.41	.18
	3	-.13	.096	.785	-.42	.17
	4	.22	.096	.279	-.08	.51
	6	.24	.095	.189	-.06	.53
6	1	-.35*	.095	.009	-.64	-.06
	3	-.36*	.096	.007	-.66	-.07
	4	-.02	.096	1.000	-.31	.28
	5	-.24	.095	.189	-.53	.06

Based on observed means. The error term is Mean Square(Error) = 1.822.

Table 29

Scheffé Testing of Topics on the Document Cohesion Construct

(I) Topic	(J) Topic	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	3	-.03	.093	.998	-.32	.25
	4	.31*	.094	.031	.02	.59
	5	.15	.092	.649	-.14	.43
	6	.43*	.093	.000	.15	.72
3	1	.03	.093	.998	-.25	.32
	4	.34*	.094	.012	.05	.63
	5	.18	.093	.449	-.11	.47
	6	.47*	.093	.000	.18	.76
4	1	-.31*	.094	.031	-.59	-.02
	3	-.34*	.094	.012	-.63	-.05
	5	-.16	.093	.569	-.45	.13
	6	.13	.094	.759	-.16	.42
5	1	-.15	.092	.649	-.43	.14
	3	-.18	.093	.449	-.47	.11
	4	.16	.093	.569	-.13	.45
	6	.29*	.093	.046	.00	.57
6	1	-.43*	.093	.000	-.72	-.15
	3	-.47*	.093	.000	-.76	-.18
	4	-.13	.094	.759	-.42	.16
	5	-.29*	.093	.046	-.57	.00

Based on observed means. The error term is Mean Square(Error) = 1.733.

Analysis of the Method Variable

Analytic method effect was also tested by Scheffé against all four of the main constructs. Results are reported in Tables 30 through 33. For each of the four constructs, the rank order of means was NMF first, LSA second, and LDA last. However, none of these comparisons are significant at the .05 level. Arguably, NMF could be considered significantly different from LDA with respect to *Term-Topic Association* and *Topic Clarity* with p-values 0.056 and 0.059 respectively. Two points are worth considering at this point. First, this is a very large sample. Large samples will naturally move toward significance. Second, even though this variable seems to be moving toward significance, it does not have explanatory power as

demonstrated in Tables 22 through 25. This means that the analytic method has no impact on the clarity of results.

Table 30

Scheffé Test Results of Method on the Term-Topic Association Construct

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.10	.070	.365	-.27	.07
	3	-.17	.070	.056	-.34	.00
2	1	.10	.070	.365	-.07	.27
	3	-.07	.070	.618	-.24	.10
3	1	.17	.070	.056	.00	.34
	2	.07	.070	.618	-.10	.24

The error term is Mean Square(Error) = 1.638.

Table 31

Scheffé Test Results of Method on the Document-Topic Association Construct

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.01	.069	.992	-.18	.16
	3	-.16	.069	.066	-.33	.01
2	1	.01	.069	.992	-.16	.18
	3	-.15	.069	.088	-.32	.02
3	1	.16	.069	.066	-.01	.33
	2	.15	.069	.088	-.02	.32

The error term is Mean Square(Error) = 1.590.

Table 32

Scheffé Test Results of Method on the Topic Clarity Construct

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.08	.074	.556	-.26	.10
	3	-.18	.074	.059	-.36	.01
2	1	.08	.074	.556	-.10	.26
	3	-.10	.074	.431	-.28	.09
3	1	.18	.074	.059	-.01	.36
	2	.10	.074	.431	-.09	.28

The error term is Mean Square(Error) = 1.822.

Table 33

Scheffé Test Results of Method on the Document Cohesiveness Construct

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.11	.072	.311	-.29	.07
	3	-.16	.072	.085	-.34	.02
2	1	.11	.072	.311	-.07	.29
	3	-.05	.072	.787	-.23	.13
3	1	.16	.072	.085	-.02	.34
	2	.05	.072	.787	-.13	.23

The error term is Mean Square(Error) = 1.733.

Interaction Term Analysis

To explore the relationships among the elements of the interaction terms, Marginal Means were collected. These values provide the mean interaction values of each of the Analysis Methods × Discussion Topic. These results are reported in tables 34 through 37 as well as Figures 12 through 15. In each construct, the highest interaction term mean value is with a combination of the LSA analytic method and discussion Topic 3. Oddly, the lowest level mean value for an interaction pair is also LSA in combination with discussion Topic 4. This interaction pairing has the lowest mean value in all four constructs.

Table 34

Term-Topic Marginal Means for Method × Topic Interaction Variable

Method		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
LDA	LDA1	5.110	.110	4.895	5.325
	LDA3	4.868	.113	4.647	5.089
	LDA4	4.938	.113	4.717	5.159
	LDA5	4.887	.111	4.670	5.105
	LDA6	4.669	.111	4.452	4.887
LSA	LSA1	5.206	.110	4.991	5.421
	LSA3	5.323	.112	5.103	5.543
	LSA4	4.559	.114	4.336	4.782
	LSA5	5.129	.111	4.910	5.347
	LSA6	4.771	.108	4.559	4.984
NMF	NMF1	5.135	.111	4.918	5.353
	NMF3	5.189	.111	4.971	5.408
	NMF4	5.107	.112	4.888	5.326
	NMF5	5.063	.107	4.853	5.274
	NMF6	4.838	.112	4.618	5.059

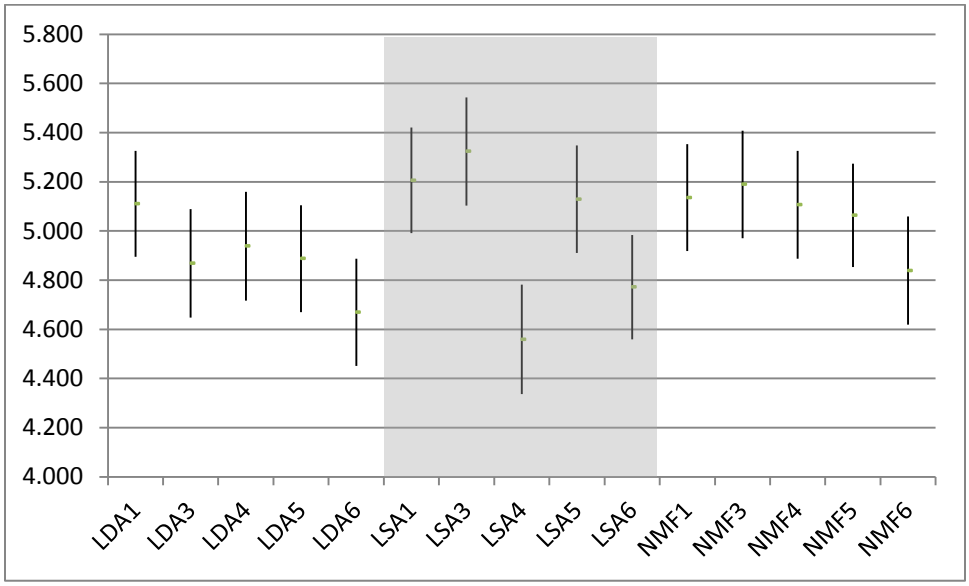

Figure 12. Term-topic marginal means confidence intervals for Method × Topic interaction.

Table 35

Document-Topic Marginal Means for Method × Topic Interaction Variable

Method		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
LDA	LDA1	5.176	.108	4.964	5.388
	LDA3	5.047	.111	4.829	5.264
	LDA4	5.116	.111	4.899	5.334
	LDA5	4.887	.109	4.673	5.102
	LDA6	4.789	.109	4.575	5.004
LSA	LSA1	5.162	.108	4.950	5.374
	LSA3	5.369	.111	5.152	5.586
	LSA4	4.488	.112	4.269	4.708
	LSA5	5.144	.110	4.929	5.359
	LSA6	4.900	.107	4.691	5.109
NMF	NMF1	5.226	.109	5.011	5.440
	NMF3	5.235	.110	5.020	5.450
	NMF4	5.130	.110	4.914	5.346
	NMF5	5.296	.106	5.088	5.503
	NMF6	4.938	.111	4.722	5.155

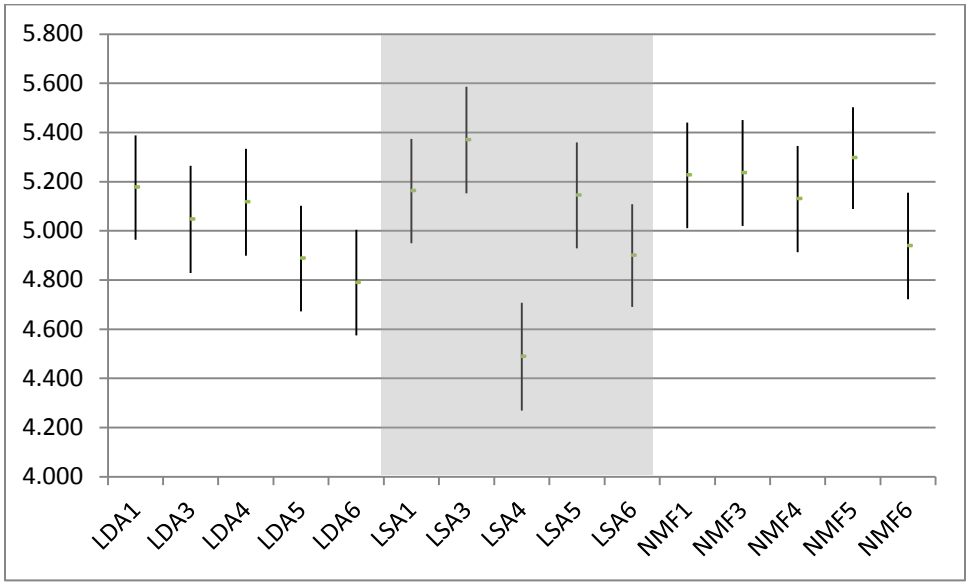

Figure 13. Document-topic marginal means confidence intervals for Method × Topic.

Table 36

Topic Clarity Marginal Means for Method × Topic Interaction Variable

Method		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
LDA	LDA1	5.110	.116	4.883	5.337
	LDA3	4.860	.119	4.627	5.094
	LDA4	4.938	.119	4.705	5.171
	LDA5	4.805	.117	4.575	5.034
	LDA6	4.692	.117	4.462	4.921
LSA	LSA1	5.103	.116	4.876	5.330
	LSA3	5.354	.118	5.122	5.586
	LSA4	4.449	.120	4.214	4.684
	LSA5	5.083	.117	4.853	5.314
	LSA6	4.829	.114	4.605	5.052
NMF	NMF1	5.173	.117	4.943	5.402
	NMF3	5.197	.117	4.967	5.427
	NMF4	4.977	.118	4.746	5.208
	NMF5	5.141	.113	4.919	5.363
	NMF6	4.815	.118	4.583	5.048

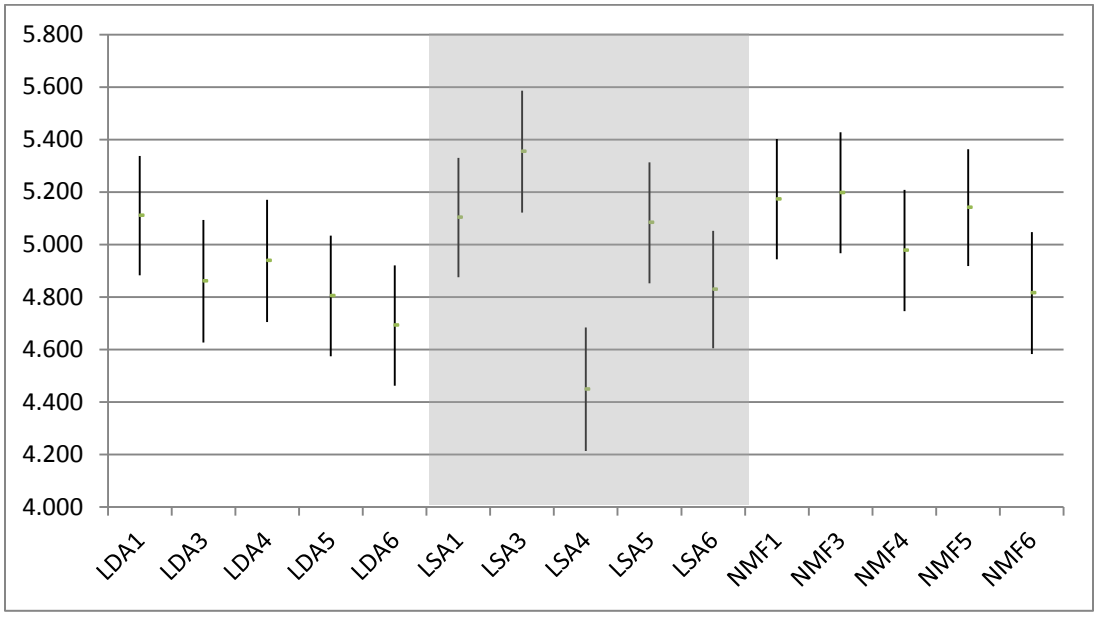

Figure 14. Topic clarity marginal means confidence intervals for Method × Topic interaction.

Table 37

Document Cohesion Marginal Means for Method × Topic Interaction Variable

Method		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
LDA	LDA1	5.118	.113	4.896	5.339
	LDA3	4.744	.116	4.517	4.971
	LDA4	4.961	.116	4.734	5.189
	LDA5	4.684	.114	4.460	4.908
	LDA6	4.586	.114	4.363	4.810
LSA	LSA1	5.051	.113	4.830	5.273
	LSA3	5.508	.115	5.281	5.734
	LSA4	4.378	.117	4.149	4.607
	LSA5	4.985	.115	4.759	5.210
	LSA6	4.741	.112	4.522	4.960
NMF	NMF1	5.083	.114	4.859	5.307
	NMF3	5.098	.115	4.874	5.323
	NMF4	4.969	.115	4.744	5.195
	NMF5	5.134	.110	4.917	5.350
	NMF6	4.615	.115	4.389	4.842

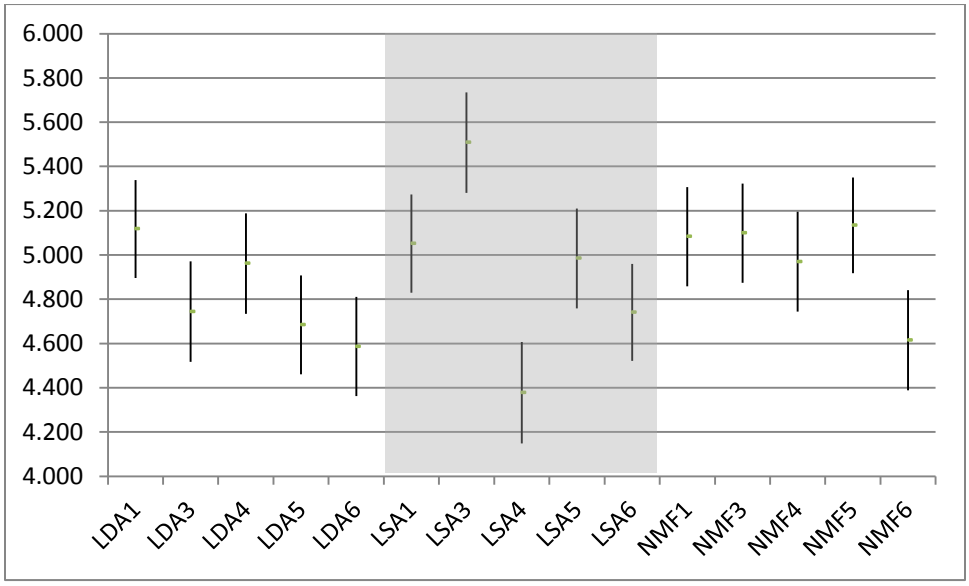

Figure 15. Document cohesion marginal means confidence intervals for Method × Topic.

Interaction Term Based on Topic Extraction Order

As mentioned in the section *Analysis of the Topic Variable*, topic numbers were assigned as topics were discovered across the analysis results. This strategy has allowed the comparison of singular discussion topics with their particular probability characteristics across analytic methods. Of further interest is the interpretability of topics given the order of extraction by analytic method. All analytic methods extract topics in order of covariance explained. That is, the first topic extracted explains the most variation in the **X**-matrix with subsequently extracted topics explaining declining amounts of variation. Figure 16 illustrates for the *Topic Clarity* construct the marginal means of the Method × Topic interaction variable with the topics sorted by extraction order within method. Generally speaking, one expects a decrease in the clarity of topics as the amount of variation explained decreases. If that were true, Figure 16 would exhibit decreasing means within each method.

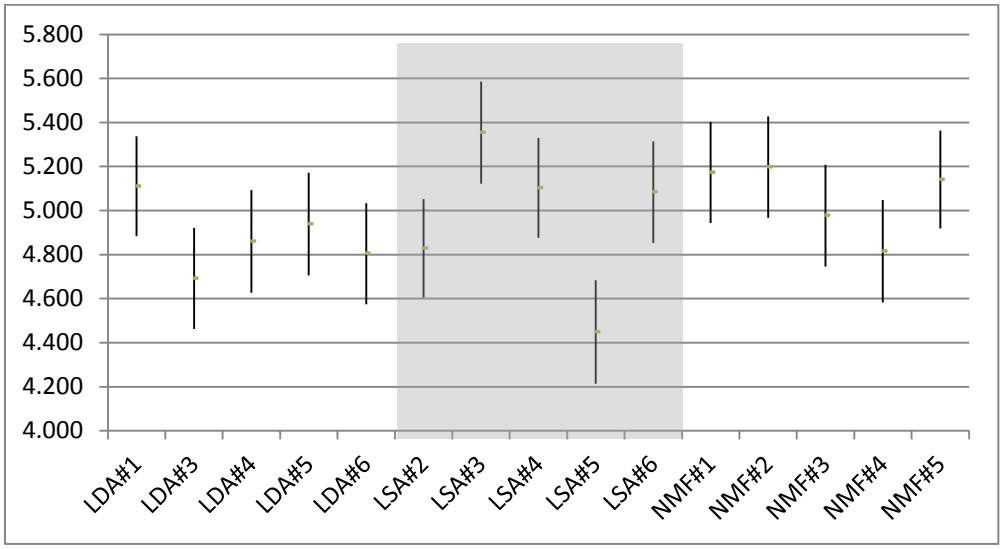

Figure 16. Marginal means confidence intervals for the Method × Topic interaction variable with the topic clarity construct sorted and labeled by extraction order.

Recall also that the documents presented for even numbered topics were extracted beginning at the 50th percentile. Since the presented documents for the even numbered topics had lower factor scores, the expectation was for lower means for Topic Clarity. That did not systematically occur with any of the analytic methods. While a logical hypothesis would be that the higher levels of variance explained increases topic clarity, this data does not support it.

While the analytic method was found to be insignificant at explaining the variation in most of the constructs, certain topics were significantly different than others. Why those topics were different was not explained by the order in which they were extracted by the analysis method. I propose that the nature of the underlying topic e.g. the singularity of the topic or uniqueness within the corpus, results in some effect on the probabilities of the **X**-matrix. What exactly generates this effect is unclear.

Step 5: Regression Analysis of the A Priori Model

With the constructs known to be clear of an order effect and know to be sound, linear regression analysis was performed of the main a priori model. The main model is Figure 7 and features *Topic Clarity* as the dependent construct and *Document-Topic Association* and *Term-Topic Association* as the predictor variables. The overall regression model is excellent with 73.8% of the variation in the dependent variable explained by the predictors (see Table 38). *F*-statistic testing of the overall model is also significant (*p*-value = 0.000) meaning that the means of the predictors are not equal to zero. The slopes of the predictors also tested successfully using the *t*-statistics (*p*-value = 0.000). Multicollinearity is not a problem in the model as measured by the variance inflation factor (VIF = 3.507). Full results are reported in Tables 38 through 40.

Table 38

A Priori Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.859[a]	.738	.738	.698

a. Predictors: (Constant), D-T Assoc, T-T Assoc

Table 39

Multiple Regression ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	2736.100	2	1368.050	2812.846	.000[b]
	Residual	968.825	1992	.486		
	Total	3704.925	1994			

a. Dependent Variable: TopicClarity
b. Predictors: (Constant), D-T Assoc, T-T Assoc

Table 40

Multiple Regression Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	.193	.066		2.939	.003		
Term-Topic Assoc	.339	.023	.321	14.981	.000	.285	3.507
Document-Topic Assoc	.610	.023	.570	26.581	.000	.285	3.507

Step 6: Regression Analysis Including Demographics

With a significant a priori base model defined, the focus is shifted to demographic information that might contribute to the explanation of a topic's clarity. During the data collection, a variety of demographic information was collected. This included information on the respondents familiarity (*FAMILIAR*) with text mining procedures. The assumption was that

few if any of the respondents would be familiar with text mining; however, if they were familiar with text mining, the results might reflect some kind of skewness as a result. Respondents were also asked if English was their first language (*EFL*). The main task respondents had to perform was interpretation of linguistic material in English. Non-native speakers might have difficulty with the task that skewed their results. The number of semester hours of language courses (*SHLC*) a respondent had completed was also collected because of the interpretive nature of the data. Potentially, higher levels of training in languages could make the respondent more sensitive to semantic context. Therefore, respondents with high levels of SHLC could have different results from those with lower levels of SHLC. How competent (*COMP*) the individual felt about categorizing the texts and how well they understood (*U*) the text passages was also collected. Here the thinking was that persons that did not feel comfortable performing the task would generate results with different means. Gender (*FEMALE*), age (*AGE*), and completed education level (*ED*) were also collected.

Analysis of the demographic data was performed by an entry and step-wise regression to determine the significant predictors of *topic clarity* construct. Some respondents did not provide all demographic variable data. The analysis was performed by deleting observations with missing data and by replacing missing values by mean values. Deleting and replacing missing values made no difference to the conclusion. The change in in variance explained was from .146 for the model with missing values deleted to .159 for the model that replaced missing values by the mean value. A significant model of demographics was found by stepwise regression that consisted of the *English as a First Language* and *Understanding* variables. This model replaces missing values for consistency in degrees of freedom and while consisting of highly significant variables it only explains 15.6% of the variation in the dependent. Also, see Tables 41 and 42

for the rest of the model. This particular regression model was unnecessary; however, it does demonstrate the importance of language and comprehension on the task.

Table 41

Demographics Model ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	576.694	2	288.347	183.614	.000
Residual	3128.232	1992	1.570		
Total	3704.925	1994			

a. Dependent Variable: TopicClarity

Table 42

Demographics Explanatory Model

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	2.644	.160		16.554	.000		
Understanding	.605	.033	.384	18.603	.000	.996	1.004
EnglishFirstLanguage	.164	.047	.071	3.460	.001	.996	1.004

a. Dependent Variable: TopicClarity

The presence of the Document-topic and Term-Topic constructs changes the significant demographic's variables. In an effort to acquire a more robust model, all variables were regressed by step-wise method resulting in a new final model. This model has an $R^2 = 0.747$ and an R^2_{adj} of 0.746. For this model, missing values for demographic variables was again replaced by mean values. Replacing missing values only improved the models variance explained by .003 and it does not change the outcome of significant variables. This is an improvement of 0.9%

variation explained from the base model with only the main constructs (see Table 38). The model's ANOVA and coefficients are documented in Tables 43 and 44. This model suggest that while the main constructs of the model are the prominent components, the greater the understanding the respondents had of the passages (coefficient = 0.084), and the more competent they felt about the task (coefficient = 0.059), the higher the explanation of the variation in the Topic Clarity construct. Likewise, older respondents scored Topic Clarity higher (coefficient = 0.105). Finally, when the respondents first language was English (coefficient = 0.066), resulted in a slightly higher Topic Clarity scores. These are intuitive results, however, the coefficients are much smaller which is consistent with demographics contributing little to the final model.

Table 43

Regression ANOVA – Main Constructs Plus Demographics

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	2767.364	6	461.227	977.984	.000
Residual	937.561	1988	.472		
Total	3704.925	1994			

a. Dependent Variable: TopicClarity

Table 44

Regression Coefficients – Main Constructs and Demographics

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.417	.108		-3.861	.000		
DocTopicAssoc	.588	.023	.550	25.675	.000	.278	3.603
TermTopicAssoc	.328	.022	.311	14.692	.000	.284	3.525
Understanding	.086	.022	.054	3.896	.000	.655	1.526
EnglishFirstLanguage	.075	.026	.033	2.837	.005	.968	1.033
AGE	.106	.034	.036	3.143	.002	.977	1.023
Competence	.063	.026	.033	2.466	.014	.733	1.365

a. Dependent Variable: TopicClarity

In an effort to acquire a more comprehensive model, all variables were regressed by entry method resulting in a new final model. Results of this model are presented in Tables 45 through 47 with significance of all variables reported. In this model, for the variable Method, LDA was set as the base model because LDA was previously identified as possessing the lowest overall mean value when predicting Topic Clarity. Topic 6 is also absorbed into the base model. In this model, missing values in the demographic variables were replaced by mean values. Table 48 presents the same model, analyzed by general linear model (Univariate analysis) adding the interaction term Method × Topic, but without replacing missing values in the demographic variables with their means. In the Univariate analysis, the interaction term is demonstrated as non-significant.

Table 45

Regression Model Summary with All Variables Included

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865[a]	.748	.746	.687

Table 46

Regression ANOVA with All Variables Included in the Model

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	2769.811	12	230.818	489.224	.000[b]
Residual	935.114	1982	.472		
Total	3704.925	1994			

a. Dependent Variable: TopicClarity

Table 47

Regression Coefficients with All Variables Included in the Model

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.438	.113		-3.864	.000		
Understanding	.087	.022	.055	3.960	.000	.654	1.528
Age	.106	.034	.036	3.140	.002	.977	1.023
EnglishFirstLanguage	.076	.026	.033	2.868	.004	.967	1.034
Competence Categorizing Text	.063	.026	.032	2.443	.015	.732	1.366
Term-Topic Assoc	.326	.022	.309	14.535	.000	.282	3.549
Document-Topic Assoc	.587	.023	.549	25.573	.000	.276	3.624
LSA	.044	.038	.015	1.166	.244	.745	1.342
NMF	.031	.038	.011	.811	.418	.745	1.343
Topic1	.042	.049	.012	.863	.388	.620	1.613
Topic3	.026	.049	.008	.541	.589	.627	1.596
Topic4	-.047	.049	-.014	-.961	.337	.632	1.582
Topic5	.004	.048	.001	.084	.933	.623	1.606

a. Dependent Variable: TopicClarity

Table 48

General Linear Model Results Including Interaction Terms

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	6.290	1	6.290	13.432	.000
	Error	815.551	1741.652	.468		
Document-Term Association	Hypothesis	284.604	1	284.604	605.758	.000
	Error	904.895	1926	.470		
Term-Topic Association	Hypothesis	100.910	1	100.910	214.778	.000
	Error	904.895	1926	.470		
Understanding	Hypothesis	7.619	1	7.619	16.216	.000
	Error	904.895	1926	.470		
CompetenceCategorizingText	Hypothesis	2.769	1	2.769	5.894	.015
	Error	904.895	1926	.470		
Age	Hypothesis	4.467	1	4.467	9.508	.002
	Error	904.895	1926	.470		
EnglishFirstLanguage	Hypothesis	3.865	1	3.865	8.226	.004
	Error	904.895	1926	.470		
Method	Hypothesis	.539	2	.269	1.113	.375
	Error	1.941	8.018	.242		
Topic	Hypothesis	1.570	4	.393	1.622	.259
	Error	1.942	8.023	.242		
Method * Topic	Hypothesis	1.935	8	.242	.515	.846
	Error	904.895	1926	.470		

Effect of Removing Outliers and Influential Observations

To assess the impact of removing the 109 outliers and influential observations (see the section *Residual Analysis* in this chapter), the final regression model with demographic data was reran using all of the 2,104 observations described at the beginning of this chapter. This analysis held the model found in the previous section steady. As previously discussed, a number of respondents did not provide full demographic data resulting in 52 observations being dropped from this analysis. The resulting data set consisted of 2,052 observations that explain 67.9% of the variation (R^2_{adj}) in the dependent variable *Topic Clarity* when regressed onto the six independent variables. The presents of the outliers and influential observations reduced the R^2_{adj} by 6.4% (.743-.679). Additionally, *English as a First Language* becomes an insignificant variable (*p*-value = .092). The full model is documented in Tables 45 through 47.

Table 49

Regression Model Summary Including Outlier and Influential Observations

R	R Square	Adjusted R Square	Std. Error of the Estimate
.825[a]	.680	.679	.800

Table 50

Regression Model ANOVA Including Outlier and Influential Observations

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	2782.112	6	463.685	723.807	.000[b]
Residual	1310.069	2045	.641		
Total	4092.181	2051			

Table 51

Regression Coefficients Including Outlier and Influential Observations

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.410	.125		-3.290	.001		
Term-Topic Assoc	.260	.021	.247	12.553	.000	.405	2.467
Doc-Topic Assoc	.621	.022	.573	28.794	.000	.395	2.533
Age	.102	.038	.034	2.692	.007	.975	1.025
Understanding	.123	.025	.076	4.984	.000	.671	1.491
Comprehension	.100	.029	.050	3.411	.001	.738	1.354
EnglishFirstLanguage	.052	.031	.021	1.685	.092	.968	1.033

If the outlier observations were allowed to remain in the data set, the standardized residuals would extend out to 4.895 standard deviations from the mean. Part of the reason for selecting ± 3 standard deviations as the cutoff for outliers is that for Z_3, only 0.27% of the observation should exceed ± 3 standard deviations. For 2,104 observations, only 5.6808 or approximately 6 observations should exceed ± 3 standard deviations. In this rerun, 34 observations exceed ± 3 standard deviations. A histogram of the residuals is available in Figure 17.

Of those 34 observations, the respondents that generated positive residuals, generally scored clarity high (5, 6, or 7), however they scored both *Document-Topic Association* and *Term-Topic Association* low (2, 3, or 4). They reported understanding the material *somewhat* or *very well* (3 or 4). Of these 34 observations with negative residuals, respondents reported Clarity very low (1, 2, or 3), however they scored both *Document-Topic Association* and *Term-Topic Association* high (4, 5, or 6). The majority of this group reported they understood the material only *somewhat*, however this group had a lot more variation in self-reported understanding than those with positive residuals with scores ranging from 1-4 out of a possible four responses. All

but two of these 34 observations were generated by respondents that self-reported as 18-30 years of age.

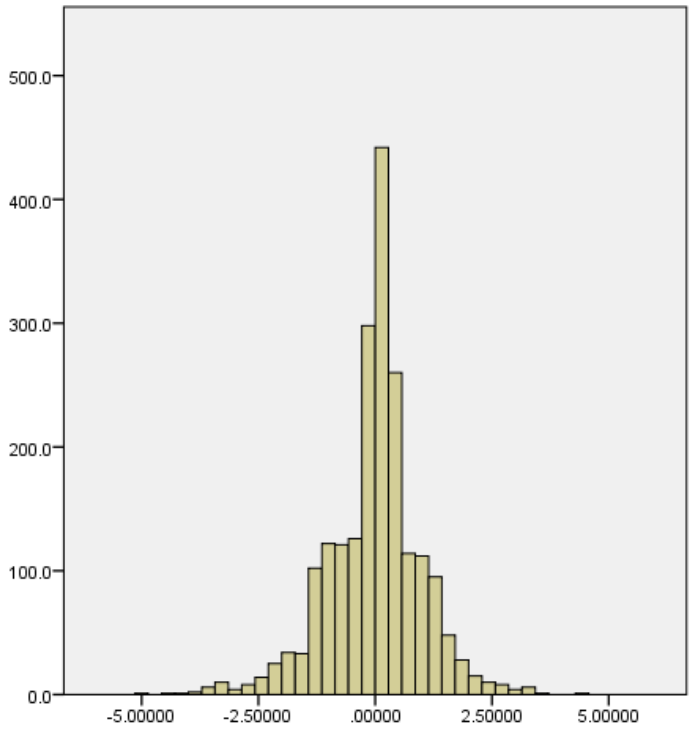

Figure 17. Histogram of standardized residuals.

With regards to influential observations, a histogram of Mahalanobis D^2 is documented in Figure 18. The range of Mahalanobis D^2 ranges from 0.4 through 56.8. Several different interpretations of influential observations using Mahalanobis distance exist. Observations with values of D^2/df greater than or equal 3 or 4, where the degrees of freedom are the number of predictor variables, is one standard for considering an observation as influential (Hair et al., 2006, p. 75). Given there were six predictor variables, applying this standard would result in observations greater than either 18 or 24 considered influential. Sixty-two observations possess D^2/df in excess of 18 and 27 observation possess D^2/df values in excess of 24.

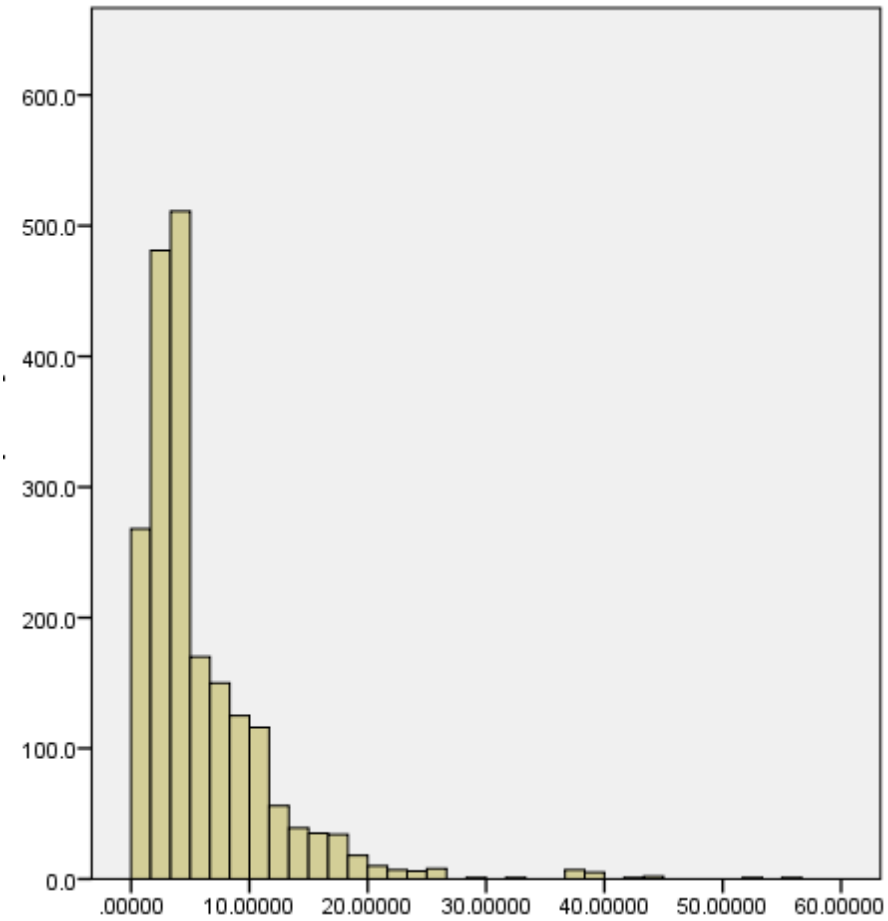

Figure 18. Histogram of Mahalanobis Distance.

An alternative to the simple D^2/df statistic is to consider D^2 values as following either a *t*-distribution (Hair, et al., 2006, p. 75) or as distributed along a χ^2 distribution (IBM , 2013), and to conservatively interpret observations with *p*-values $\leq .001$ or .005 (Hair et, al., 2006 p. 75; IBM, 2013) as influential. Describing the distribution presented in Figure 18 as a *t*-distribution is problematic at best. Since D^2/df cannot exist as a negative number, and given the general shape of the distribution in Figure 18, I choose to use the χ^2 distribution. A histogram of *p*-values for Mahalanobis distance computed as distributed along the $\chi 2$ distribution (1-

cdf.chisq(x_i,6) with six degrees of freedom is provided in Figure 19. Using the χ^2 distribution and the more conservative .001 standard, there are 33 observations to consider as influential.

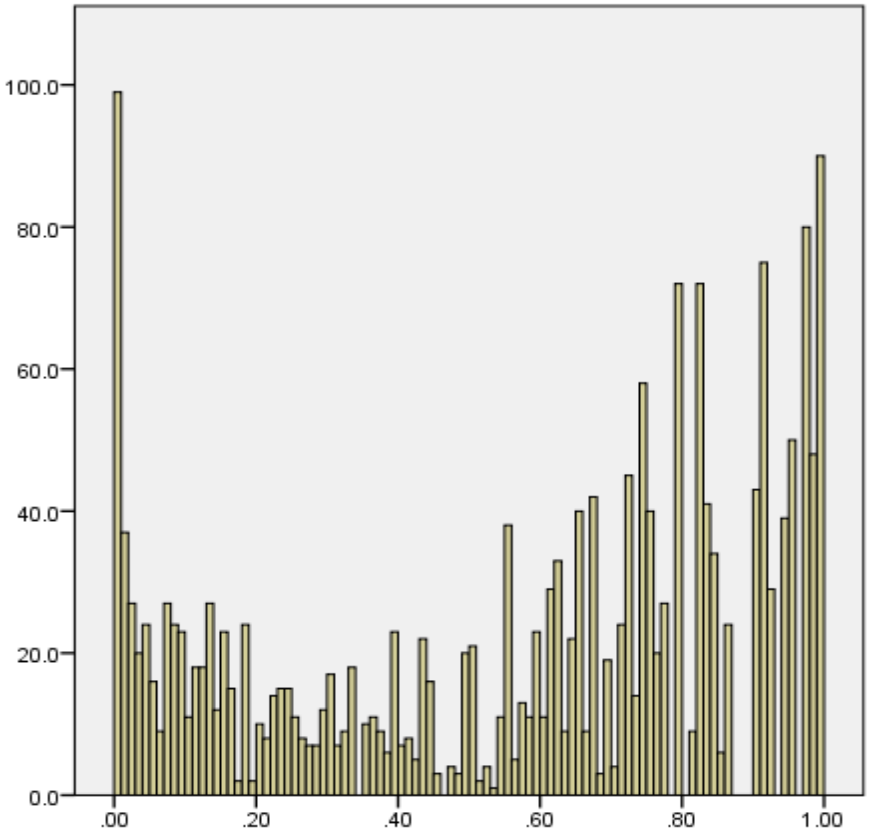

Figure 19. Histogram of p-values for Mahalanobis D distributed as a χ2 statistic

Of these 33 observations, the respondents scored Topic Clarity, Document-Topic Association, and Term-Topic Association predominantly in a range of 4-7. Additionally, many of the respondents that generated these observations said they had some familiarity with text-mining. One-third of this group reported they were still learning the English language. All reported they understood the material very well and felt relatively competent at categorizing text material. In start contract to those identified as outliers by Z-scores, this group was older with 24 of the 33 self-reported as 40 years of age or older.

While there are 67 observations identified by this regression as potential outliers or influential observations, they were generated by only 36 respondents. Further, respondent's 50, 102, and 259 generated 7 outlier or influential observations each or 21 out of 67 outliers.

Given the inconsistency of scoring in these 67 observations between the Topic Clarity construct and the Document-Topic and Term-Topic Association constructs; the oddity in the relationship between young respondents and outlier verses older respondents and influential observations; and the fact that removing these observations does not unduly contribute to the model (6.4%), deleting these observations is a reasonable alternative. Note also that retention of these outliers and influential observations are not particularly detrimental; however, they do exert considerable pressure on the interpretation of residual normality.

CHAPTER 6

DISCUSSION, CONTRIBUTION, AND FUTURE RESEARCH

Discussion

During the last two plus decades, text analytics methods have advanced rapidly. Prior to this research, each successive algorithm was demonstrated as a superior method at its introduction by using some measurement metric that was linked in some way to the algorithm design. This linkage is logical; the designer perceives a specific need or weakness in existing tools then designs a new and better mousetrap; the designer then tests the new mousetrap's ability to fill the need or correct for the perceived weakness. This rational, while logical, builds in a bias via the test measurement metric. Independent capability testing is a solution to this built in bias; however, such testing is limited and it is also based on some measurement metric.

One goal of this experiment's design was to develop a testing process that was more objective for a business user. That objectivity starts by asking, "What defines a better text mining method?" The solution to this question depends on the viewer's perspective. The algorithm designer is trying to fill a perceived or defined need while the business analyst is trying to answer a business question. To answer that business question, the analytic tool has to generate results that are clear, understandable, and interpretable. That is, the method must extract topics that are understandable and possesses relationships between the topics and terms as well as topics and documents that are clear. Since historical testing has been oriented toward the designers perspective, this user perception also represents a gap in the literature.

From this experiment, we now know that generally, the analytic method is not a significant predictor of the clarity and understanding of the extracted topic. It is also not significant when measuring the association of documents to the topic, or to the understanding of

the term and topic association. Only when I include all of the outliers in the data is there any significance in the method. However, while there is significance in the variable β, the model does not have explanatory power possessing R^2 values ranging from 0.02 through 0.04 depending on the construct it is regressed with. While not significant, the rank order of means in all four constructs was LDA < LSA < NMF. This is exactly backwards from the pilot testing conducted in summer 2012.

These results were unexpected. During informal pilot testing, the analytic method was significant with LDA out performing LSA in a study involving only those two methods. In a second informal pilot, the analytic method was again found significant with the order of means NMF < LSA < LDA. It is unclear what caused this complete reversal of results. Additionally, during those informal studies, the range of means across topics by LSA was much tighter than the other analytic methods. In this study, Figures 14 through 15 document the marginal means and confidence intervals of each topic as extracted by each method. In these Figures, it is clear that LSA had the broadest range of means.

The first pilot study, considering only LDA and LSA, and used a set of data generated by queries ran against Twitter. These queries were diversely structured around three distinctly different topic domains including *healthcare*, *Olympics 2012*, and *economics*. Each query was further subdivided into three additional subgroups generating a corpus consisting of nine unique and diverse topics. This was a heterogeneous data set. This data was generated based on non-conformities in service quality at a Pizza restaurant (Anaya, 2011, p. 64). The specific causes for non-conformities in this data was controlled and could extend from one of four major causes originating because of either *food condition*, *food processing methods*, *facilities*, or *customer service*. Each of the major causes was further broken down into five or six subcategories.

During this data collection, respondents were creatively writing about experiences that may or may not have occurred, originated over an expansive time frame, and did not possess an anchoring factor e.g. specific restaurant franchise. This data, while possessing looseness because of a lack of origin singularity, was less diverse than the data used in the first pilot study. The main experiment was conducted using data collected by a business entity attempting to answer a singular and specific question: why are customers closing their accounts. The collected data set were customer responses received during one business day. These customer expressions originated in response to a common experience. While a homogeneity scale for text data does not exist, the characteristics of this data set are by far more homogeneous. This seems to suggest that diverse topic structures, heterogeneous corpora's, are analyzed more effectively by LDA while homogeneous corpora's are analyzed more effectively by NMF. This issue requires further investigation.

Extracted discussion topics are also significant predictors of user perceived topic clarity. This could very well be an extension of the corpora homogeneity. While a pattern seems to emerge from the order number assigned to the topics, those numbers are not descriptive of the order in which the data was extracted from the corpus. The numbering associated with the topics reflects the order in which topics that spanned all three analytic methods were discovered during visual screening of algorithm results.

Additionally, the interaction term Method × Topic was also a significant predictor of user perceived clarity (see Tables 22-25). This means there exists some unidentified characteristic in the corpus that interacts differently with each of the algorithms. However, this result only holds true while considering each of the constructs in isolation. When Method × Topic interaction are considered in the larger more comprehensive model, it is not significant (see Table 48).

From the regression model, the Document-Topic relationship (b = .717) is a stronger predictor of the Topic Clarity than is the Term-Topic Association (b = .242). The overall regression model is very strong explaining 73% of the variation in Topic Clarity. While several of the demographics variables were significant as predictors, they could have been ignored since they only explain 0.7% of the variation in Topic Clarity. The Understanding and Competence variables produced somewhat surprising results. The concepts these variables are measuring are so close that Multicollinearity was expected. However, as measured by VIF, Multicollinearity does not exist in any of the variables.

During data topic extraction, LDA was noted having difficulty in extracting topics from the more homogeneous corpuses. This difficulty was exhibited in its inability to extract documents and terms with high probabilities. This led to the algorithm being run multiple times with each subsequent iteration executed with ever-increasing algorithm parameters e.g. increase number of random starts, reduced convergence tolerance, and increased number of iterations. NMF likewise had difficulty with the data set used in this dissertation taking 78.6 minutes to generate a valid solution while LDA required 0.75 minutes and LSA required 0.02 minutes. In separate research, LDA typically takes much longer to process than it did with this particular data. When applied against the first pilot study data set, heterogeneous Twitter data, LDA performed much faster than LSA and extracted topics successfully using no random restarts, 500 iterations with a convergence tolerance of 10^{-6} for variational inference and 1,000 iterations and a tolerance of 10^{-4} for parameter estimation. These are the default configuration parameters for the R software package Topic Models. Since both NMF and LDA use optimization routines, there is more of an artistic component needed to arrive at a successful solution. While not the

main focus of this research, there is a need for guidance among the algorithms based on homogeneity of the target corpus.

Contribution

This dissertation contributes to the body of work in four important ways. First, it offers an experimental designed method of testing algorithms. In this experiment, potential users rate the methods based on their perception of clarity of the results. This is a more logical method of testing the algorithm outputs since a mixture of usability and clarity of output will ultimately drive the user's intent to reuse the algorithm, increases their confidence in the analytic results, and facilitates decision making in the business process.

The introduction of an experimentally designed testing process for text mining algorithms removes or at least reduces the biased measurement systems used in the past. The procedure introduced here should become the standard across all bias prone algorithm testing. Eliminating bias sources is critical for determining the most methods of analysis.

Second, this dissertation develops a new model under which to test text-mining systems. This model extends the framework started in Chang et al. (2009). The third contribution of this dissertation extends from that model and is a new instrument for testing the clarity of text mining methods.

Finally, this dissertation discovers that from a user's perspective, the analytic method is not a significant predictor of topic clarity, term-topic association, or document-topic association. This discovery reaffirms that of Kintsch and Mangalath (2011). With such information, the business practitioner as well as the researcher can focus on other characteristics such as price and ease of use when making plans to deploy new analytic resources.

Future Research

As a result of this research, new research has already begun into various elements of the corpus structure and how those elements impact algorithm solutions. Corpus components under consideration in this new research include the *number of words that define a topic*, *topic semantic singularity*, and *corpus topic homogeneity*. The number of words that are extracted as defining a corpus topic appears descriptive of the cohesiveness or semantic singularity of a topic. In this research, NMF defined topics with a range of word counts running from 7 to 21. Likewise LSA extracted topics had a word count range of 6 to 22 and LDA word count range ran from 6 to 18. For the topic identified as number 1 in the analysis presented in chapter 4, NMF extracted that topic with only 7 words, while LSA extracted it based on 12 words. Likewise, the discussion topic identified as topic 6 in the analysis was extracted by NMF in 9 words, LSA in 13 words, and by LDA in 6 words. An extracted topic structure could consist of a simple singular or narrowly defined discussing topic *products arrive late*, complex *products arrive late because mail service is slow*, or compound *products arrive late and you never have what I want when I want it. I prefer the competition*. Defining word counts appear to be indicative of singularity, which seems to lead to cohesiveness of a topics structure. If this relationship holds, then new metrics can be explored that measure the *Consistency of Term-Topic* construct directly.

Homogeneity of a corpus is important in selecting the most time efficient algorithm as well as the most convenient to operate. Unfortunately, as of right now, corpus homogeneity is a very subjective and there is not any agreed upon strategy for testing it. There may however, be some ratio's that can be introduced that will measure homogeneity. For example, the ration of *count of word with sparsity less than X* over *count of raw words* in the corpus might suffice to measure homogeneity. A homogeneous corpus should discuss closely related topics and use

more or less the same words in doing so. In such a ratio, a homogeneous corpus should score high giving the analyst/researcher some kind of guidance in algorithm selection.

Existing text mining literature does not examine corpora structures. The closest any of the literature gets to discussing this discovery are references that discuss implications of document length within the LSA literature. These references are particularly concerned with the **X**-matrix weighting schemes (Sultan and Buckley, 1988; Evangelopoulos et al., 2012). The understanding that corpus homogeneity affects an algorithm's effectiveness is an important discovery. However, this understanding raises many questions and has implications for the researcher as well as the practitioner communities. Homogeneity of the data results from a lack of randomness in the sample. The researcher in text mining algorithms should expect many business corpora to exhibit homogeneity meaning they should become more versatile in algorithm usage.

APPENDIX A

INSTITUTIONAL REVIEW BOARD

APPLICATION NO. 12-494

UNT

A green light to greatness.™

Office of the Vice President of Research and Economic Development
OFFICE OF RESEARCH SERVICES

November 5, 2012

Dr. Nick Evangelopoulos
Student Investigator: Triss Ashton
Department of Information Technology and Decision Science
University of North Texas

RE: Human Subjects Application No. 12-494

Dear Dr. Evangelopoulos:

In accordance with 45 CFR Part 46 Section 46.101, your study titled "Accuracy and Interpretability Testing of Tex Mining" has been determined to qualify for an exemption from further review by the UNT Institutional Review Board (IRB).

No changes may be made to your study's procedures or forms without prior written approval from the UNT IRB. Please contact Jordan Harmon, Research Compliance Analyst, ext. 3940, if you wish to make any such changes. Any changes to your procedures or forms after 3 years will require completion of a new IRB application.

We wish you success with your study.

Sincerely,

P Kaminski, PhD

Patricia L. Kaminski, Ph.D.
Associate Professor
Chair, Institutional Review Board

PK:jh

UNIVERSITY OF NORTH TEXAS
1155 Union Circle #305250 Denton, Texas 76203-5017
940.565.3940 940.565.4277 fax http://research.unt.edu

Minimal Review Application

For IRB Use Only	
File Number:	
Approval:	

University of North Texas Institutional Review Board
OHRP Federalwide Assurance: FWA00007479

Section I: Filling Out and Saving the Form

Please use Adobe Acrobat to fill out this form and submit it along with all supplemental documents to the IRB Office as described in the Electronic Submission Checklist on page 6.

If you do not have Adobe Acrobat, please use Adobe Reader to fill out the form. To save your changes, go to "File" in the top tool bar and choose "Save As...PDF." To save future changes, follow the same steps but replace your existing document with the current changes.

For Mac users, right click the web link of the application you would like to open. Click "download linked file." After the file has downloaded, drag the file from your bottom tool bar to your desktop to save. Right click the file and click "Open with" and select Adobe Acrobat Pro.

Section II: Does this Form Apply?
Please check the box indicating your answer to each of the following questions:

1. Will your research study involve any vulnerable populations such as children, prisoners, pregnant women or mentally disabled persons?	YES ☐	NO ☑
2. Could public disclosure of any identifiable data you collect place the participants at risk of criminal or civil liability or be damaging to the participants' financial standing, employability or reputation?	YES ☐	NO ☑
3. Will your study involve data collection procedures other than surveys, educational tests, interviews, or observation of public behavior?	YES ☐	NO ☑
4. Will your study involve any sensitive subject matters such as: abortion, criminal activity, sexual activity, sexually transmitted diseases, prior diagnosis for mental health disorders, or victims of violence?	YES ☐	NO ☑
5. Will your study involve audio-recording or video-recording the participants?	YES ☐	NO ☑
6. Will your study involve obtaining individually identifiable information from health care plans, health care clearinghouses, or health care providers?	YES ☐	NO ☑

If you answered YES to any of the above questions, your study will not meet the criteria for Minimal Review. Please fill out the Expedited or Full Board Application for your study.

Section III: General Information
Type only in the blue fields, and closely follow all stated length limits. Handwritten forms will not be accepted.

1. Title of Study (Must be identical to the title of any related internal or external grant proposal)

ACCURACY AND INTERPRETABILITY TESTING OF TEXT MINING METHODS

2. Investigator (or Supervising Investigator for Student Studies)
Must be a full-time UNT faculty member or a full-time UNT staff employee whose job responsibilities include conducting human subjects research. A faculty **Supervising Investigator** is required for all student studies which require IRB review, including theses and dissertations. Student Investigator information is entered in Section 4.

First Name	Last Name	Email Address
Nick	Evangelopoulos	Nick.Evangelopoulos@unt.edu

UNT Department	UNT Building & Room Number	Office Phone Number
COB-ITDS	BLB-365D	940-565-3056

3. Co-Investigator (if applicable)

First Name	Last Name	Email Address

UNT Department or University	Title

4. Student Investigator (if applicable, for student studies such as theses and dissertations)

First Name	Last Name	Email Address
Triss	Ashton	Triss.Ashton@unt.edu

UNT Department	Degree Program
COB-ITDS	PhD Management Science

5. Key Personnel
(List the name of all other Key Personnel (including students) who are responsible for the design, conduct, or reporting of the study (including recruitment or data collection).

N/A

NIH or CITI IRB Training
Have you, any Co-Investigator, any Student Investigator, and all Key Personnel completed the required NIH IRB training course ("Protecting Human Research Participants") or the CITI IRB training course ("Human Subjects Research") and electronically submitted a copy of the completion certificate to untirb@unt.edu? YES ☑ NO ☐

If you answered "No," this training is required for all Key Personnel before your study can be approved. The NIH IRB course may be accessed by visiting: http://phrp.nihtraining.com. The CITI IRB course may be accessed by visiting: https://www.citiprogram.org/.

6. Funding Information (if applicable)
Has funding been proposed or awarded for this project? YES ☐ NO ☑
If yes, please submit the statement of work or a project summary and provide the proposal number or project ID number for any external funding or the account number for any internal funding for this project.

7. Purpose of Study

In the space provided, briefly state the purpose of your study in lay language, including the research question(s) you intend to answer. A brief summary of what you write here should be included in the Informed Consent Form.

Mixtures of discussion topics exist in collections of textual communications that are written by business customers. Discovering those topics and grouping the communications by topic is important for decision making and is accomplished by a group of mathematical techniques called text mining. Many text mining techniques exist, however, which one is best for any particular situation is unknown. More specifically, how users perceive the results of the various text mining techniques is unknown.

The purpose of this study is to test the perceived effectiveness and accuracy of text mining methods. The specific research question this study attempts to answer the research question:

"Which algorithms do humans perceive as more effective at extracting topics and classifying documents to topics?"

8. Recruitment of Participants

Describe the projected number of subjects.

This study will involve approximately 300 subjects recruited from sections of DSCI 2710 in the College of Business.

Describe the population from which subjects will be recruited (including gender, racial/ethnic composition, and age range).

The participant demographic mix is consistent with the enrollment of the recruited DSCI 2710 class. The composition of these classes are mixed gender with slightly more male than female, mixed ethnic composition but predominantly Caucasian, and from 19 to approximately 60 years in age. Generally the demographics of the subjects will mirror that of the UNT campus.

Describe how you will recruit the subjects.

Participation will be voluntary. The principal investigator will announce the study, its purpose, and the studies procedure in class and provide an internet link to the subjects that wish to participate via Blackboard Learn and if necessary by e-mail.

There are no other recruitment materials.

Have you attached a copy of all recruitment materials including flyers, e-mails, and scripts for classroom announcements?

YES ☑ NO ☐

9. Location of Study

Identify all locations where the study will be conducted.

The study will only be conducted at UNT. The instrument is offered over the internet and subjects can complete it at any location they desire.

For data collection sites other than UNT, have you attached a signed and dated letter on the cooperating institution's letterhead giving approval for data collection at that site?

YES ☐ NO ☐

10. Informed Consent

Describe the steps for obtaining the subjects' informed consent (by whom, where, when, etc.).

The study is conducted by internet survey. The first step of the instrument is the presentation of the informed consent notice. At the bottom of the informed consent notice, the subject is offered the option to agree to participate. The survey will only advance if the subject clicks on "Yes". The text as presented is:

"I have read, understood, and printed a copy of, the above consent form and desire of my own free will to participate in this study."

Yes
No

11. Informed Consent Forms

Written Informed Consent Forms to be signed by the subject after IRB approval are required for most research projects with human participants (exceptions include telephone surveys, internet surveys, and other circumstances where the subject is not present; an Informed Consent Notice may be substituted). Templates for creating informed consent forms and notices are located on the IRB website at **http://research.unt.edu/faculty-resources/research-integrity-and-compliance/use-of-humans-in-research**. **Final drafts of all informed consent documents you plan to use must be submitted before IRB review can begin.**

12. Foreign Languages

Will your study involve the use of any language other than English for Informed Consent forms, data collection instruments, or recruitment materials?

YES ☐ NO ☑

If "Yes," after the IRB has notified you of the approval of the English version of your forms, you must then submit the foreign language versions along with a back-translation for each. Specify **all** foreign languages below:

13. Data Collection

Which methods will you use to collect data?

- ☐ Interviews
- ☐ Surveys
- ☐ Focus Groups
- ☑ Internet Surveys
- ☐ Review of Existing Records
- ☐ Observation
- ☐ Other (Please list below)

If "Review of Existing Records" and/or "Observation" are checked above, please describe below the records you plan to review and/or the observations you plan to make for your study.

N/A

Have you attached a copy of all data collection instruments and interview scripts to be used?

YES ☑ NO ☐

What is the estimated time for a subject's participation in each study activity (including time per session and total number of sessions)?

One session lasting 45-60 minutes depending of the subjects reading and decision making speed.

14. Compensation

Describe any compensation subjects will receive for participating in the study. Include the timing for payment and any conditions for receipt of such compensation. If extra credit for a course is offered, an alternative non-research activity with equivalent time and effort must also be offered.

At the discretion of the instructor, the student participants will receive extra academic credit. As an alternative, non-participants in the class may complete two Hawkes Learning Systems assignments that are complementary to the course's scheduled study at the time of this study is administered.

15. Risks and Benefits

Describe any foreseeable risks to subjects presented by the proposed study and the precautions you will take to minimize such risks.

There are no foreseeable risks associated with this research beyond possible anxiety resulting from making judgements.

Describe the anticipated benefits to subjects or others (including your field of study).

The subjects will receive no benefit beyond exposure to text mining results. The results will help others direct future research and make informed decision about which text mining method to implement.

16. Confidentiality

Describe the procedures you will use to maintain the confidentiality of any personally identifiable data.

No personal identifiable data will be collected beyond name and participants class for the purpose of awarding extra credit. The data collected will be stored in the HIPPA compliant Qualtrics secure database until it has been deleted by the primary investigator. Working data for analysis will be anonymous because any personal identification information will be removed. The confidentiality of the individual subject will be maintained in any publications or presentations generated out of this data.

Please specify where your research records will be maintained on the UNT campus, any coding or other steps you will take to separate participants' names/identities from research data, and how long you will retain personally identifiable data in your research records. Federal IRB regulations require that the investigator's research records (including signed Informed Consent forms) be maintained for at least 3 years following the end of the study.

Anonymous working data will be maintained in room BLB-367A and the College of Business computer systems. The instrument is an internet based survey and signed consent forms will not be collected.

17. Publication of Results

Please identify all methods in which you may publicly disseminate the results of your study.

- [x] Academic Journal
- [ ] Academic conference paper or public poster session
- [ ] Book or chapter
- [x] A thesis or dissertation for one of your students
- [ ] UNT Scholarly Works Repository
- [ ] Other (Please list below; e.g. Website or blog)

Investigator or Supervising Investigator Certification

- [x] By checking this box and e-mailing this application to the UNT IRB **from my UNT e-mail account**, I am certifying that the information in this application is complete and accurate. I agree that this study will be conducted in accordance with the UNT IRB Guidelines and the study procedures and forms approved by the UNT IRB.

Electronic Submission Checklist

1. Attach all supplementary documents, including:
    a. Copies of all NIH or CITI IRB Training completion certificates not previously submitted to the IRB Office;
    b. A copy of the statement of work or project summary for any internal or external funding for this study;
    c. A copy of all recruitment materials;
    d. A copy of the approval letter from each data collection site (other than UNT);
    e. A copy of all Informed Consent forms or notices; and
    f. A copy of all data collection instruments, interview scripts and intervention protocols.

2. The application and all supplementary documents must be **e-mailed from the Investigator's or Supervising Investigator's UNT e-mail account to** untirb@unt.edu. Please insert **"Minimal Review"** in the subject line of your e-mail.

Contact **Jordan Harmon** at Jordan.Harmon@unt.edu for any questions about completion of your application.

APPENDIX B

EXPERIMENT INSTRUMENT

Introduction

The instrument used for data collection during experiment 2 is a very long and repetitive instrument. The instrument was generated utilizing the results of text mining analysis performed on data set 3. This data set had six consisted of six discussion topics. These topics were extracted by each of three text mining methods. This resulted in a total of 18 possible results that prospective users could evaluate. All 18 were assembled in combination with a Likert scale instrument. When a respondent entered the survey instrument, Qualtrics randomly selected 9 of the 18 possible results and presented them for evaluation by the respondent. The instrument also consisted of an informed consent statement, and demographic data collection items.

Since the total instrument is so long (44 pages) only a mask of the key components including the informed consent, study introduction, a sample of the presented data, Likert instrument, and demographics items are presented here. Additionally, a discussion of and comparison of the results extracted by the text mining are presented.

University of North Texas Institutional Review Board

Informed Consent Notice

Before agreeing to participate in this research study, it is important that you read and understand the following explanation of the purpose, benefits and risks of the study and how it will be conducted.

Title of Study: ACCURACY AND INTERPRETABILITY TESTING OF TEXT MINING METHODS

Student Investigator: Triss Ashton, University of North Texas (UNT) Department of ITDS. Supervising Investigator: Nick Evangelopoulos

Purpose of the Study: You are being asked to participate in a research study which involves testing the effectiveness and accuracy of text mining methods. Text mining starts with a collection of documents. The objective of a text mining method is to identify the underlying topics, prioritize the topics according to frequency of occurrence, and then associate the various documents to these topics. The present study compares three existing text mining methods. You are being asked to participate in a research study which involves testing the effectiveness and accuracy of text mining methods.

Study Procedures: You will be asked to review nine sets of results that were generated by a text mining method. After reviewing each set of results you will be asked questions concerning your opinion of the results. This study will take about 45 minutes to one-hour of your time.

Foreseeable Risks: No foreseeable risks are involved in this study with the possible exception of anxiety related to decision making.

Benefits to the Subjects or Others: This study is not expected to be of any direct benefit to you, but we hope to learn more about interpreting text mining output. The results will help others direct future research and make informed decision about which text mining method to implement.

Compensation for Participants: At the discretion of the instructor, the student participants will receive extra academic credit. As an alternative, those not wishing to participate in the study may complete two Hawkes Learning Systems assignments for the same level of extra credit. The two assignments will be complementary to the course's scheduled instruction topics at the time of this study is administered.

Procedures for Maintaining Confidentiality of Research Records: The data collected will be stored in the HIPPA compliant Qualtrics secure database until it has been deleted by the primary investigator. The confidentiality of your individual information will be maintained in any publications or presentations regarding this study.

Questions about the Study: If you have any questions about the study, you may contact Triss Ashton at 940-369-8379 or by email at Triss.Ashton@unt.edu or Nick Evangelopoulos at 940-565-3056 or by email at Nick.Evangelopoulos@unt.edu.

Review for the Protection of Participants: This research study has been reviewed and approved by the UNT Institutional Review Board (IRB). The UNT IRB can be contacted at (940) 565-3940 with any questions regarding the rights of research subjects.

Research Participants' Rights: Your participation in the survey confirms that you have read all of the above and that you agree to all of the following:

- Triss Ashton has explained the study to you and you have had an opportunity to contact him/her with any questions about the study. You have been informed of the possible benefits and the potential risks of the study.
- You understand that you do not have to take part in this study, and your refusal to participate or your decision to withdraw will involve no penalty or loss of rights or benefits. The study personnel may choose to stop your participation at any time.
- Your decision whether to participate or to withdraw from the study will have no effect on your grade or standing in this course.
- You understand why the study is being conducted and how it will be performed.
- You understand your rights as a research participant and you voluntarily consent to participate in this study.
- You understand you may print a copy of this form for your records.

Study Introduction

The purpose of this survey is to evaluate the effectiveness of three text-mining tools. Text mining starts with a collection of documents. In business, these documents are normally related, for example, customer comments that complaint about service, or a collection of SEC filings. The collection will have a series of recurring topics running through them that are repeatedly discussed in several documents. The objective of a text mining method is to identify the recurring topics, prioritize the topics according to frequency of occurrence, and then associate the various documents to these topics. Search engines, like Google, are based on some form of text mining methodology.

In the presentation, a set of documents will be presented to you as they were received and analyzed. That means there will be spelling, punctuation, and grammatical errors. The errors are retained in the data because it is unrealistic to edit documents in an operational environment where thousands of documents are involved, therefore, the analytic method must be capable of handling such discrepancies.

When an analytic result is presented to you, it will start with a list of keywords and their relative importance. Keywords with high relative importance are more important in defining the discussion topic. Review the keywords noting their relative importance, review the documents, and then answer the questions that follow.

Sample of Presented Data

Treatment 1: Method A, Topic 1 (This form is repeated for each of 18 treatments**)**

A text analysis technique revealed a discussion topic based on the following frequently occurring keywords (listed with relative importance):

KEYWORDS

ship	0.26		vendor2	0.09
product1	0.22		queue	0.04
faster	0.20		alway	0.02
day	0.16			

Here are ten documents that are supposed to be related to the topic:

- ship the products at the top of my queue that show 'available' and only show them 'available' when you can ship them. That really annoyed me when products down near the bottom of my queue arrived when the first 5 product1s on my queue all showed available.
- Ship product1s faster. It seemed to take almost 5 days to get a product1 that was available now. Get a closer distribution center for faster turn around of products. I loved the free coupons and that would want me to choose BB in the future over the competition
- Too many product1s have a wait. Get more copies of these. And ship them when available. I'd rather have a product1 at the top of my queue, even if it takes an extra day or two to get here coming from a further distribution center. Time for product1s to arrive
- Ship the product1s shown as 'Available Now' in order, starting from the top of the queue. vendor2 never ships selecions #5 and #8 from my queue because the shipping lead time is two days or longer.
- find a faster way of shipping. and when custumer calls in product1s that were shipped on his behalf as not being recieved by vendor1 on time, vendor1 should just ship out custumers products in que. it takes too long for product1s to arrive to custumer. An
- Get the product1s to the customer faster. The turn around was to slow. I would watch the product1s the same day and ship them back that same day or the next and it still would take 6 or 7 days or more.
- Unfortunately the vendor2 shipping facility is in my city, so their product1s arrive the next business day. vendor1's distribution facility was located in the large city less than 10 miles away, but still took 2-3 business days.

- Faster shipping. vendor2 ships to me a lot faster (with a 5 product1 account I get 8-10 product1s per week.)
- Send product1s at the top of my queue, Distribution center, do not send out faster enough. I would receive e-mail stating that vendor1 has received my products; it would be two to three days, before I would get an email telling the next were shipped.

Q11 Considering only the relationship of the keyword and the discussion topic:

	Strongly Disagree (1)	Disagree (2)	Somewhat Disagree (3)	Neither Agree nor Disagree (4)	Somewhat Agree (5)	Agree (6)	Strongly Agree (7)
The keywords accurately reflect the topic being discussed.	❍	❍	❍	❍	❍	❍	❍
After examining the keywords, it was easy to understand what the topic was about.	❍	❍	❍	❍	❍	❍	❍
The keywords are helpful in understanding the topic.	❍	❍	❍	❍	❍	❍	❍
The keywords define a single topic.	❍	❍	❍	❍	❍	❍	❍
The keywords are related to each other.	❍	❍	❍	❍	❍	❍	❍

Q12 Considering only the relationship of the documents and the discussion topic:

	Strongly Disagree (1)	Disagree (2)	Somewhat Disagree (3)	Neither Agree nor Disagree (4)	Somewhat Agree (5)	Agree (6)	Strongly Agree (7)
The documents accurately reflect the topic being discussed.	❍	❍	❍	❍	❍	❍	❍
After examining the documents, it was easy to understand what the topic was about.	❍	❍	❍	❍	❍	❍	❍
The documents are helpful in understanding the topic.	❍	❍	❍	❍	❍	❍	❍
The documents define a single topic.	❍	❍	❍	❍	❍	❍	❍
The documents are related to each other.	❍	❍	❍	❍	❍	❍	❍

Q13 Considering only the discussion topic:

	Strongly Disagree (1)	Disagree (2)	Somewhat Disagree (3)	Neither Agree nor Disagree (4)	Somewhat Agree (5)	Agree (6)	Strongly Agree (7)
The discussion topic is clear to me.	❍	❍	❍	❍	❍	❍	❍
The concept of the topic is clear to me.	❍	❍	❍	❍	❍	❍	❍
I feel I know what this topic is about.	❍	❍	❍	❍	❍	❍	❍
I would be able to label this topic.	❍	❍	❍	❍	❍	❍	❍

Q14 In the box below, label (describe) the topic:

Q15 How easy was it to label (describe) this topic?

- ❍ Very Difficult (1)
- ❍ Difficult (2)
- ❍ Somewhat Difficult (3)
- ❍ Neutral (4)
- ❍ Somewhat Easy (5)
- ❍ Easy (6)
- ❍ Very Easy (7)

Q16 Consider the relationship of the documents to one another:

	Strongly Disagree (1)	Disagree (2)	Somewhat Disagree (3)	Neither Agree nor Disagree (4)	Somewhat Agree (5)	Agree (6)	Strongly Agree (7)
All of these documents talk about the same thing.	❍	❍	❍	❍	❍	❍	❍
The documents address a topic in a consistent manner.	❍	❍	❍	❍	❍	❍	❍
It would be easy to identify documents that "do not belong" in the group.	❍	❍	❍	❍	❍	❍	❍
These documents are similar to each other.	❍	❍	❍	❍	❍	❍	❍

Demographics Items

Q100 Please use the box below to provide any comments you would like to make regarding the text analysis results you looked at today.

Q101 Before this survey, how familiar were you with text mining tools, methods, and algorithms?

- None (1)
- Little (2)
- Some (3)
- Alot (4)

Q102 Is English your first language?

- Yes (1)
- No (but I am fluent) (2)
- No (I am still learning) (3)

Q103 How many semester hours of college level composition, writing, or literature have you completed (any language e.g. English, German, or Spanish)?

- Select number of hours (1)
- 3 (2)
- 6 (3)
- 9 (4)
- 12 (5)
- 15 (6)
- 18 (7)
- More than 18 (8)

Q104 How well did you understand the documents presented in this survey?

- ❍ Not at all (1)
- ❍ Not very well (2)
- ❍ somewhat (3)
- ❍ Very well (4)
- ❍ Extremely well (5)

Q105 How competent do you consider yourself at categorizing text into topic classes?

- ❍ Very competent (I read a document and understand the broader topic it discussed) (1)
- ❍ Relatively competent (I categorized documents with only occasional difficulty) (2)
- ❍ Relatively challenged (I frequently had difficulty associating with a topic) (3)
- ❍ Very challenged (When I read a document I cannot understand the topic it discussed) (4)

Q106 What is your gender?

- ❍ Female (1)
- ❍ Male (2)

Q107 What is your age category?

- ❍ 18-20 (1)
- ❍ 21-25 (2)
- ❍ 26-30 (3)
- ❍ 31-35 (4)
- ❍ 36-45 (5)
- ❍ 46-55 (6)
- ❍ 56-65 (7)
- ❍ 66-75 (8)
- ❍ 76 and above (9)

Q108 What is the highest level of education you have completed?

- ❍ Some high school or less (1)
- ❍ High School graduate (2)
- ❍ Some College education (3)
- ❍ College Freshman (4)
- ❍ College Sophomore (5)
- ❍ College Junior (6)
- ❍ College Senior (7)
- ❍ Bachelor's Degree (8)
- ❍ Some graduate education (9)
- ❍ Masters Student or Degree (10)
- ❍ Doctoral Student or Degree (11)
- ❍ Other level of qualification, please state below: (12)

Q111 For extra credit, please click the next button to link you to another page. You will be asked to provide your instructors name as well as your name. NOTE: Your name will be used only for purposes of sending your extra credit points to your instructor. Your name will not be associated with your responses in any other way. Subsequently, your name will be deleted from this database and your responses will be processed by the researchers anonymously.

Q112 What course are you taking and who is your instructor?

- ❍ Select your Instructor (1)
- ❍ DSCI 2710.002 - Ashton (2)
- ❍ DSCI 2710.004 - George (3)
- ❍ Other (4)

Q113 Put your name in the box (Last, First):

REFERENCE LIST

Allen, H., Gearan, P, & Rexer, K. (2011). *5th annual data mining survey – 2011 survey summary report*. Retrieved from: http://rexeranalytics.com/Data-Miner-Survey-Results-2011.html

Allen, H., Gearan, P, & Rexer, K. (2010). *4th annual data mining survey – 2010 survey summary report*. Retrieved from: http://rexeranalytics.com/Data-Miner-Survey-Results-2010.html

Anaya, L. (2011). Comparing latent Dirichlet allocation and latent semantic analysis as classifiers. . (Doctorial Dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3529209)

Badea, L. (2008). Extracting gene expressions profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. *Pacific symposium on Biocomputing*, *13*, 279-290.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77-85.

Bartlett, M.S. (1951). A further note on test of significance in factor analysis. *British Journal of Psychology*, *4*, 1-2.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning*, *3*, 993-1022.

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Proceeding of Advances in Neural Information Processing Systems,* Cambridge, MA, (pp. 17-24).

Blei, D. & Lafferty, J. (2007). A correlated topic model of *Science. The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. & McAuliffe, J. (2007). Supervised topic models. *Proceedings of Advances in Neural Information Processing Systems*. Vancouver, B.C., Canada, (pp. 1280-1288).

Blei, D. & Lafferty, J. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning*, Pittsburgh, PA, (pp. 113-120).

Blei, D. (2011). Introduction to probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

.

Bradford, R. (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley, CA, (pp. 153-162).

Brown, J. (2009a). Choosing the Right Type of Rotation in PCA and EFA, *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(3), 20-25.

Brown, J. D. (2009b). Statistics Corner. Questions and answers about language testing statistics: Choosing the right number of components or factors in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(2), 19-23.

Brunet,J., Tamayo, P., Golub, T., & Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceeding of the National Academy of Science, 101*, (pp. 4164–4169).

Burke, S. (2001). Missing values, outliers, robust statistics and non-parametric methods. *Statistics and Data Analysis LC-GC Europe Online Supplement, 59,* 19–24.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Durbin, J. & Watson, G., (1950). Testing for serial correlation in least squares regression: I. *Biometrika, 37*(3/4), 409-428.

Casella, G. & Berger, R. (2001). Hypothesis testing in statistics. In: N. Smelser, & P. Baltes, (Eds), *International Encyclopedia of the Social &Behavioral Sciences*, (pp. 7118-7121).

Cattell, R.B. (1966) The scree test for the number of factors. *Multivariate behavior Research, 1,* 245-276.

Chang, J., Boyd-Grabber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of Advances in Neural Information Processing Systems*. Paper presented at the Neural Information Processing Systems Foundation conference, Vancouver, B.C., Canada, (pp.288-297).

Chang, J. (2010). Not-so-latent Dirichlet allocation: Collapsed Gibbs sampling using human judgments. *Proceeding of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, (pp. 131-138).

Cliff, N. (1988). The Eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103*(2), 276-279.

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS computational biology, 4*(7), 1-12.

Ding, C., Li, T., & Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis, 52*, (3913-3927).

Ding, C., Li, T., & Peng, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence, Chi-square statistic, and a hybrid method. *Proceeding of the twenty-first national conference on artificial intelligence*, Boston, MA, (pp. 342-347).

Dumais, S., Furnas, G., Laddauer, T., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceeding of the SIGCHI conference on human factors in computing systems*. Washington, D.C., (pp. 281-285).

Efron, Miles (2005). Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science, 56*(9), 969–988.

Enis, B., Cox, K., & Stafford, J. (1972) Students as subjects in consumer behavior experiments. *Journal of Marketing Research, 9*(1), 72-74.

Evangelopoulos, N., Zhang, X., & Prybutok, V. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems, 21*, 70-86.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Feinerer, I. (2012). Package 'tm' [Software]. Available from http://cran.r-project.org/web/packages/lsa/lsa.pdf

Frigyesi, A. & Hoglund, M. (2008). Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. *Cancer Informatics, 6*, 275-292.

Gaujoux, R. & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics, 11*(367), 1-9.

Gaujoux, R. (2010). Using the Package NMF. Retrieved from http://cran.r-project.org/web/packages/NMF/vignettes/NMF-vignette.pdf

Gaujoux, R. (2012). Package 'NMF' [Software]. Available from http://cran.r-project.org/web/packages/NMF/NMF.pdf

Gaussier, E. & Goutte, C. (2005). Relation between pLSA and NMF and implications. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, (pp. 601-602).

Gentry, J. (2012). Package 'twitteR' [Software]. Available from http://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Glorfeld, L. (1995). An improvement on Horn's parallel analysis method for selecting the correct number of factors to retain. *Educational and Psychological Measurements, 55*, 377-393.

Griffiths, T. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5228-5235.

Grun, B. & Hornik, K. (2010). Topic models in R. Retrieved from http://cran.uvigo.es/web/packages/topicmodels/vignettes/topicmodels.pdf

Grun, B. & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Grun, B. & Hornik, K. (2012). Package 'topicmodels' [Software]. Available from http://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

Guttman, L., (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19* (2), 149-161.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006) *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hofmann, T. (1999a). Probabilistic Latent Semantic Indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*. Berkeley, CA, (pp. 50-57).

Hofmann, T. (1999b). Probabilistic latent semantic analysis. *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, (pp. 289–296).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177-196.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–186.

Hutchins, L., Murphy, S., Singh, P., & Graber, J. (2008). Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, *24*(23), 2684-2690.

Jackson, D.A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, *74*(8), 2204-2214.

Jaspersoft (2011). Seven Trends that will change Business Intelligence as we know it. Retrieved from: http://www.jaspersoft.com/sites/default/files/Jaspersoft%20eBook.pdf

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, Ann Arbor, MI, (pp. 29–36).

Kakkonen, T., Myller, N., Sutinen, E. (2006). Applying latent Dirichlet allocation to automatic essay grading. *Proceedings of the 5th International Conference on Natural Language Processing*, Turku, Finland. (pp. 110–120).

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, *11*(3), 275-288.

Kim, J. & Mueller, C. (1978). *Factor analysis: Statistical methods and practical issues*. Newbury Park, CA: Sage University Press.

Kim, P. & Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, *13*, 1706-1718.

Kim, Y.S., Chang, J.H., & Zhang, B.T. (2002). A comparative evaluation of data driven models in translation selection of machine translation. *Proceeding COLING '02 Proceedings of the 19th international conference on Computational linguistics*, (pp. 1-7).

Kintsch, W. & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, *3*, 346-370.

Kitchenham, B. (1996). Evaluating software engineering methods and tool Part 1: The evaluation context and evaluation methods. *ACM SIGSOFT Software Engineering Notes*, *21*(1), 11-14.

Kummamura, K., Lotlikar, R., Roy, S., Singal, K., & Krishnapuram, R. (2004). A hierarchical monothetic document clustering algorithm for summarization and browsing search results. *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, (pp. 658-665).

Laundauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Lee, D. & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788-791.

Lee, S., Song, J., & Kim, Y. (2010). An empirical comparison of four text mining methods. *The Journal of Computer Information Systems*, *51*(1), 1-10.

Leverson, I., Danielsen, A., Wold, B., & Samdal, O. (2012). What they want and what they get: self-reported motives, perceived competence, and relatedness in adolescent leisure activities. *Child Development Research*, *2012*, 1-11.

Li, W. & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlation. *Proceeding ICML '06 Proceedings of the 23rd international conference on machine learning*, Pittsburgh, PA, (pp. 577-584).

Manning, C., Raghavan, P., & Schutze, H. (2009). *An introduction to information retrieval.* Retrieved from: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Miner, G., Delen, D., Elder, J.; Fast, A., Hill, T., & Nisbit, R. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* Waltham, MA: Academic Press/Elsevier.

NIST/SEMATECH (2012). *e-Handbook of Statistical Methods*. Retrieved from: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm.

Pauca, V., Shahnaz, F., Berry, M., & Plemmons, R. (2004). Text mining using non-negative matrix factorization. *Proceedings of the fourth SIAM international conference on data mining*. Lake Buena Vista, FL. (pp. 452-460).

Pascual-Montano, A., Carazo, J., Kochi, K., Lehman, D., & Pascual-Marqui, R. (2006). Nonsmooth nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(3), (pp. 403–415).

Paatero, P. & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*, 111-126.

Peres-Neto, P., Jackson, D., & Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*, 974-997.

Peterson, R. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of consumer research*, *28*(3), 450-461.

Phan, X., Nguyen, L., & Horiguchi, S. (2008). Learning to classify short and sparce text & web with hidden topics from large-scale data collections. *Proceedings of the 17th International World Wide Web Conference (www 2008)*, Beijing, China (pp. 91-100).

Porchea, S. (2008). *The Influence of Calculator use, Familiarity and Learning Experience on a High-Stakes Examination*. (Doctorial Dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3321429)

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed Gibbs Sampling for latent Dirichlet allocation. *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, Las Vegas, NV, (pp. 567-577).

Porter, M. (1980). An algorithm for suffix stripping. *Program: Automated library and information systems*, *14*(3), 130-137.

Porter, M. (2001). Snowball: A Language for Stemming Algorithms. Retrieved from: http://snowball.tartarus.org/texts/introduction.html.

R Development Core Team (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from: http://www.R-project.org.

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.

Riordan, B. & Jones, M. (2011). Redundancy in perceptual and linguistic experience: comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3*, 303-345.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report*. Fourth Quarter 2011. Reston, VA: The Data Warehouse Institute. Retrieved from: http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

Salton , G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing, *Communications of the ACM , 18*(11), 613-620.

Shahnaz, F., Berry, M., Pauca, V., & Plemmons, R. (2006). Document clustering using nonnegative matrix factorization. *Information processing and management, 42*, 373-386.

Sidorova, A., Evangelopoulos, N., Valacich, J., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. *MIS Quarterly*, *32*(3), 467-A20.

Sirkin, R. (2006). *Statistics for the Social Sciences* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Sulton, G. & Buckley, C., (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluations methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Canada, (pp. 1105-1112).

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

Wild, F. (2007). An LSA package for R. *Proceeding of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning* (LSA-TEL'07) Heerlen, Netherlands, (pp. 11-12).

Wild, F. (2012). Package 'lsa' [Software]. Available from http://cran.r-project.org/web/packages/lsa/lsa.pdf

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, Toronto, ON, Canada, (pp. 267–273).

Zhao, Y. & Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis. *Technical Report #01-40 University of Minnesota, Department of Computer Science.* Retrieved from: https://wwws.cs.umn.edu/tech_reports_upload/tr2001/01-040.pdf

Zhao, Y. & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*(3), 311-331.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, *51*, 918-930.