Hierarchical Text Topic Modeling with Applications in Social Media-Enabled Cyber Maintenance Decision Analysis and Quality Hypothesis Generation

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Zhenhuan Sui

Graduate Program in Industrial and Systems Engineering

The Ohio State University

2017

Dissertation Committee:

Professor Theodore T. Allen, Advisor

Professor Steven Neil MacEachern

Professor Cathy Honghui Xia

Professor Nena L. Couch

ABSTRACT

Many decision problems are set in changing environments. For example, determining the optimal investment in cyber maintenance depends on whether there is evidence of an unusual vulnerability such as “Heartbleed” that is causing an especially high rate of incidents. This gives rise to the need for timely information to update decision models so that the optimal policies can be generated for each decision period. Social media provides a streaming source of relevant information, but that information needs to be efficiently transformed into numbers to enable the needed updates. This dissertation first explores the use of social media as an observation source for timely decision-making. To efficiently generate the observations for Bayesian updates, the dissertation proposes a novel computational method to fit an existing clustering model, called K-means Latent Dirichlet Allocation (KLDA). The method is illustrated using a cyber security problem related to changing maintenance policies during periods of elevated risk. Also, the dissertation studies four text corpora with 100 replications and show that KLDA is associated with significantly reduced computational times and more consistent model accuracy compared with collapsed Gibbs sampling.

Because social media is becoming more popular, researchers have begun applying text analytics models and tools to extract information from these social media platforms. Many of the text analytics models are based on Latent Dirichlet Allocation (LDA). But

these models are often poor estimators of topic proportions for emerging topics. Therefore, the second part of dissertation proposes a visual summarizing technique based on topic models, a point system, and Twitter feeds to support passive summarizing and sensemaking. The associated "importance score" point system is intended to mitigate the weakness of topic models. The proposed method is called TWitter Importance Score Topic (TWIST) summarizing method. TWIST employs the topic proportion outputs of tweets and assigns importance points to present trending topics. TWIST generates a chart showing the important and trending topics that are discussed over a given time period. The dissertation illustrates the methodology using two cyber-security field case study examples.

Finally, the dissertation proposes a general framework to teach the engineers and practitioners how to work with text data. As an extension of Exploratory Data Analysis (EDA) in quality improvement problems, Exploratory Text Data Analysis (ETDA) implements text as the input data and the goal is to extract useful information from the text inputs for exploration of potential problems and causal effects. This part of the dissertation presents a practical framework for ETDA in the quality improvement projects with four major steps of ETDA: pre-processing text data, text data processing and display, salient feature identification, and salient feature interpretation. For this purpose, various case studies are presented alongside the major steps and tried to discuss these steps with various visualization techniques available in ETDA.

To my beloved family

Bo Dai, Zhongxue Sui, and other family and host family members

For their endless love and support along the way

To my advisor

Professor Theodore T. Allen

Without his guidance, inspiration and encouragement

This dissertation couldn't been done

ACKNOWLEDGMENTS

I could not complete this dissertation without the assistance and support of many people. First, my most sincere appreciation goes to my PhD advisor Professor Theodore T. Allen, for his generous guidance and advice during my PhD study and research in this dissertation. As my advisor and mentor, he has always been nice, patient, inspiring and supportive. No words could express my gratitude and thank towards him for being my PhD advisor. I truly admire his work, knowledge and academic research passion. I would also thank Mrs. Jodie Allen for proofreading our papers. I am also deeply thankful for Major Nathan Parker and the TRADOC Analysis Center in the U.S. Army for funding part of my research (W9124N-15-T-0033) and the text analytics software development opportunity to work with Major Parker and the U.S. Army. NSF grant #1409214 also partially supported my research. It is truly my honor to serve the TRADOC Analysis Center in the U.S. Army to deliver software used in four organizations in U.S. Army, although I do not own the access to sensitive data of U.S. Army. I am also grateful to Professor Steven Neil MacEachern, Professor Mathew Roberts, and Professor Cathy Honghui Xia (ordered by last names) and Professor Nena L. Couch for kindly being in my committee of Candidacy Exam, Colloquium, and PhD defense, and providing extremely valuable insights, feedback and suggestions. Many thanks to my previous advisors, Professor Suvrajeet Sen, Professor Joseph Fiksel, Professor Shahrukh Irani,

Professor Marc E. Posner, and Professor Nicholas G. Hall for introducing me to the wonders of scientific research and everything they have done for me. I would also like to thank Professor Jerald Brevick, Professor Antonio Conejo, Professor Ramteen Sioshansi, Professor Philip Smith, Professor Carolyn Sommerich, and Professor Julia Higle for teaching me, letting me in the PhD program at OSU and funding my PhD study. Without all the professors' inspiration and support, I would not start my PhD study in The Ohio State University.

I would like to thank Kaveh Akbari for his joint work on Chapter 4 and proofreading for the dissertation. I would also like to thank my other friends Professor Anthony Afful-Dadzie, Dr. Chengjun CJ Hou, Dr. Chen Xie, Dr. Shijie Huang, Dr. Sayak RoyChowdhury, Dr. Yue Tan and other numerous friends in the United States and China for their friendship and help during my PhD study. I would like to thank Mrs. Helena Law and Dr. Harold Law for giving me the scholarship for my undergraduate study at Washington University. Without their generous help, I could finish my undergraduate and PhD studies. Also, my thanks go to my previous teachers and professors, and my internship bosses Dr. Chad Farschman at Owens Corning, Dr. Mingjun Zhao and Dr. Santosh Mishra at State Street, Mr. Colleen Benson and Dr. Gavin Duffy at Goldman Sachs Tokyo, and my upcoming bosses Mr. Tariq Javed and Mr. Dean Tavakoli for their teaching, guidance, and help that leads to my job career in the field I have passions in for the past and future.

Special thanks to Jiaoyue Wang for the encouragement, help, support and love during my last year of PhD study. Also, special thanks to upcoming Dr. Simeng Karen Li

for the time 2004-2015. Last, but not least, I would like to thank my parents Bo Dai, Zhongxue Sui, my other beloved family members, and host family members (Dr. Rebecca Stilson, Dr. Mike Sullivan, Rachel Sullivan, Hannah Sullivan, and Sam Sullivan) for their unconditional love, encouragement and support over years.

VITA

2011..... B.A. Physics, Minor in Mathematics
Denison University, USA

2011..... B.S. Systems Science & Engineering
Minor in General Business
Washington University in St. Louis, USA

2011 -2014 ..... Graduate Teaching Associate,
Department of Integrated System Engineering
The Ohio State University, USA

2012 - 2017 ..... Graduate Research Associate,
Department of Integrated System Engineering
The Ohio State University, USA

2014..... Data Scientist Intern, Modeling &
Optimization Team, Research & Development
Owens Corning, USA

2015..... Quantitative Associate Intern
Enterprise Risk Analytics
State Street, USA

2016..... Desk Quantitative Strategist Intern
Goldman Sachs, Japan

FIELDS OF STUDY

Major Field: Industrial and Systems Engineering

Specialization: Operations Research

Minor Fields: Statistics, Econometrics

PUBLICATIONS

Allen, T. T.; Sui, Z.; and Parker, N. (under 2nd review). "Timely Decision Analysis Enabled by Efficient Social Media Modeling". Submitted to *Decision Analysis*.

Allen, T. T.; Sui, Z.; and Akbari, K. (under 1st review). "Exploratory Text Data Analysis for Quality Hypothesis Generation". Submitted to *Journal of Quality Technology*.

Allen, T. T.; Parker, N.; and Sui, Z. (2016). "Using Innovative Text Analytics on a Military Specific Corpus". Abstract. *Military Operations Research Conference*, Quantico, VA, June.

Sui, Z. and Allen, T. T. (2016). "NLP, LDA, SMERT, k-Means and Efficient Estimation Methods with Military Applications". Abstract. *INFORMS 2016*, Nashville, November.

Sui, Z.; Milam, D.; and Allen T.T. (2015). "A visual summarizing technique based on importance score and Twitter feeds". *INFORMS Social Media Analytics Student Paper Competition. Finalist.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Decision Making with Text Modeling Motivation

The information security is a booming market. According to Gartner, worldwide organizations spent $76.9 billion in 2015 on information security. In 2015, the average total cost of losing sensitive corporate or personal information is approximately $3.8 Billion. Therefore, cyber investment decision-making has received considerable attention (Paté-Cornell, 2012; Gao, Zhong, Mei S, 2013; Parnell, Butler, Wichmann, Tedeschi, Merritt, 2015; Miller, Wagner, Aickelin, Garibaldi, 2016).

As social media is becoming more popular, researchers have begun applying text analytics models and tools to extract information from social media platforms. Social media data has been studied with the goal of increasing participation in public policy decision-making (Charalabidis and Loukis, 2012). Others have investigated decisions about social media selections by individuals (Bok et al., 2012). Yet, how can we leverage social media data to support decision analyses unrelated to the social media? Traditional estimation methods in topic models include collapsed Gibbs sampling and variational inference methods. Yet, how can these methods be computationally efficient for corpora

involving tens of thousands of documents? How to make the results of these estimation methods repeatable and stable?

1.2 Quality Engineering with Text Modeling Motivation

Many relevant human events can be viewed as samples from multinomial distributions. For example, a word in speech could be a sample from a topic distribution or a cyber security system could shift from one loss category to another. In many cases, it is important to model and understand the probabilities of these events through soliciting human inputs or through observations. A widely cited method for developing intuitive clusters in free style text is Latent Dirichlet Allocation (LDA), which includes multinomial samples for both words and topics or clusters associated with each word. Allen, Xiong, and Afful-Dadzie (2016) propose subject matter expert refined topic (SMERT) for probabilistic clustering of texts to permit experts or users to edit the topics using knowledge about the system or their own needs. SMERT and LDA estimate the proportion of words in the overall corpus on each topic. As a special case of LDA, SMERT potentially incorporates “high-level” inputs from a subject matter expert to adjust the topics and clusters by zapping or boosting words in the topic definitions. But there are issues with SMERT estimation including that the collapsed Gibbs sampling is potentially unacceptably noisy and the topic proportion estimates can be biased or “shrunk” toward the mean. But, how could we leverage the models to identify an emerging topic? Moreover, how will quality engineers and decision analysts deal with text data, visualization and analysis?

1.3 Dissertation Overview

Unstructured text data in surveys, social media, and ordinary media is occupying an increasing space. Quality engineers and decision analysts can benefit but they need tools to apply their formulations, concepts, and perspectives on text data to quality improvement and decision making problems. Therefore, the dissertation will give the solutions to these issues. This dissertation has five chapters. The first is this introduction which describes the challenges in cyber security decision making using Twitter and text analytics visualization. Chapter 2 addresses the major problem statements which are:

1. How can we leverage social media data to support decision analyses unrelated to the social media?

2. Traditional estimation methods in topic models include collapsed Gibbs sampling and variational inference methods. Yet, how can these methods be computationally efficient for corpora involving tens of thousands of documents? How to make the results of these estimation methods repeatable and stable?

In Chapter 3, a visual summarizing technique based on topic models is proposed with the support of passive summarizing and sensemaking from Twitter feeds. It is mainly to addresses the problem:

3. How could we leverage the models to identify an emerging topic?

In Chapter 4, a general framework on how to work with text data to generate quality hypothesis will be discussed. It is mainly to addresses the problem:

4. How will quality engineers and decision analysts deal with text data, visualization and analysis?

The final chapter summarizes the findings and proposes future researches of other topic modeling estimation methods and the application of topic modeling on decision making problems in the financial industry.

Here, Chapters 2 is joint work with Major Nathan Parker from the TRADOC Analysis Center in the U.S. Army, and Chapter 3 is joint work with David Milam, Chapter 4 is joint work with Kaveh Akbari. Professor Theodore Allen is the co-author for all the chapters.

1.4 References

Allen, T.T.; Xiong, H.; and Afful-Dadzie A. (2016). "A Directed Topic Model Applied to Call Center Improvement". *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Blei, D.M.; Ng, A.Y.; and Jordan, M.I. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3, pp. 993-1022.

Bok, H.S.; Kankanhalli, A.; Raman, K.S.; and Sambamurthy, V. (2012). "Revisiting Media Choice: a Behavioral Decision-Making Perspective". *International Journal of e-Collaboration (IJeC)* 8(3), pp. 19-35.

Charalabidis, Y. and Loukis, E. (2012). "Participative Public Policy Making through Multiple Social Media Platforms Utilization". *International Journal of Electronic Government Research (IJEGR)* 8(3), pp. 78-97.

Gao, X.; Zhong, W.; and Mei, S. (2013). "Information Security Investment When Hackers Disseminate Knowledge". *Decision Analysis* 10(4), pp. 352-368.

Gartner (2014). "Gartner Says Worldwide Information Security Spending Will Grow Almost 8 Percent in 2014 as Organizations Become More Threat-Aware". Retrieved from http://www.gartner.com/newsroom/id/2828722 Accessed November 22, 2016.

Miller, S.; Wagner, C.; Aickelin, U.; and Garibaldi, J.M. (2016). "Modelling Cyber-Security Experts' Decision Making Processes Using Aggregation Operators". *Computers & Security* 62, pp. 229-245.

Parnell, G.S.; Butler III, R.E.; Wichmann, S.J.; Tedeschi, M.; and Merritt, D. (2015). "Air Force Cyberspace Investment Analysis". *Decision Analysis*. 12(2), pp. 81-95.

Paté-Cornell, M.E. (2012). "Games, Risks, and Analytics: Several Illustrative Cases Involving National Security and Management Situations". *Decision Analysis* 9(2), pp. 186-203.

Sui, Z.; Milam, D.; and Allen, T.T. (2015). "A Visual Summarizing Technique Based On Importance Score and Twitter Feeds". *INFORMS Social Media Analytics Student Paper Competition*.

CHAPTER 2
LATENT DIRICHLET ALLOCATION TOPIC MODELING AND TIMELY DECISION ANALYSIS ENABLED BY EFFICIENT SOCIAL MEDIA MODELING

2.1 Introduction

Consider solving a similar problem for several decision periods. In each period, an updated model is used. Decisions in different periods are assumed here to have no interactive effect and time enters as a factor only through changes to the environment for different periods. As a result, a solution must differ from those involving optimizing sequential decision policies such as decision trees (e.g., Cao, (2014)) or continuous control policies (e.g., Borrero et al. (2015)). The motivating example is a cyber security investment decision (Paté-Cornell (2012); Gao et al. (2013); Parnell et al. (2015); Miller et al. (2016)). The motivating problem considered here relates to monthly basic maintenance for periods of usual or, alternatively, elevated risk. We assume that attackers are making their decisions on much faster time scales than months. Therefore, we ignore both the game theoretic and sequencing aspects and focus on updating only.

The primary objective of this article is to provide computationally efficient and repeatable methods for updating period-specific decision models using Twitter or other streaming text data such as Facebook. Instead of using social media as an application area

for decision analyses (Charalabidis and Loukis (2012), Bok et al., (2012)), we seek to use it as a source of timely data for decision analysis problems. In a few periods in our motivating example vulnerabilities constituted a large fraction of cyber security expert Tweets. These were precisely the periods in which the most warnings and incidents were eventually observed. By summarizing Tweets, it was possible for system administrators to anticipate and mitigate the attacks that followed. We expect that social media-based summarizing could aid updating for many types of decision analysis problems.

The method must be computationally efficient because the needed processing is often prohibitively slow (e.g., see Blei et al. (2003); Packiam and Prakash (2015)). For our purposes, any method that transforms streaming text into numbers that strongly correlate with the system state could serve. For example, one might use sentiment analysis which scores positive and negative words or even simpler counts of word mentions. Here, we use clustering methods primarily because the Twitter experts wrote about many irrelevant subjects. Through clustering, all their topics can be mapped including those that relate to the decision problem. By recalling a small number of Tweets primarily in the key clusters, data can be generated.

Probably the most widely studied methods for clustering text data are variants of Latent Dirichlet Allocation or "topic models" (Blei et al. (2003); Packiam and Prakash (2015)). There are several ways to fit topic models to data including collapsed Gibbs sampling, a form of Markov chain Monte Carlo simulation (Teh et al. (2006)), and "mean field variational inference" (Blei et al. (2003)), an approximate maximum likelihood fit of the clustering (distribution) model.

Yet, both collapsed Gibbs sampling and variational inference can be prohibitively expensive computationally for corpora involving tens of thousands of documents. Collapsed Gibbs is known for its lack of repeatability. Here, we seek computationally efficient methods to fit approximate topic models with improved repeatability. Specifically, we propose to explore the concept of transforming k-means clustering results to estimate topic model parameters. Lee (2012) had used fuzzy c clustering to generate "fuzzy LDA" which permits documents that cover multiple topics like LDA and unlike k-means clustering. Yet, Ghosh and Dubey (2013) show that k-means is O(T) and fuzzy c clustering is order O(T^2), where T is the number of clusters. Therefore, we seek an O(T) method that permits fractional membership.

The remainder of this chapter is organized as follows. First, we describe the time decision analysis formulation. Then, we describe the proposed methods for efficient clustering needed to generate the decision formulation inputs. Next, we illustrate the methods on a cyber investment problem. Finally, we compare the proposed estimation methods with alternatives and conclude with a summary of the implementation and future work possibilities.

2.2 Timely Decision Modeling

Consider a two-phase approach for timely decision-making. The first phase is a startup phase in which the model is estimated and matrices are estimated to facilitate Bayesian updates. The second phase is steady state in which new text data are analyzed and Bayesian updates potentially change the results for subsequent decision problems.

The Bayesian updates require the collection of observation data and the estimation of "observation matrices" (Smallwood and Sondik (1973)), both of which steps we describe.

2.2.1 Two Phase Approach

In the first phase, the decision analysis problem is formatted. We denote the system state as Y with possible values y =1, …, s and the chosen action in period i is a_i = 1, …, a. The reward as it depends on the action and state is $r[(y, a_i)]$ and the utility function is $u[r(y, a_i)]$. The current probability distribution is $p_i(y, a_i)$ and the initial probability distribution is $p_0(y, a_0)$. In each time period, the decision-maker selects the option which maximizes the expected utility, ρ, given by:

$$\max_{a_i} \rho(a_i) = \sum_{y=1}^{s} p_i(y, a_i) u[r(y, a_i)] \quad (2.1)$$

which is essentially the Von Neumann and Morgenstern (2007) problem. Note that the utility could be equivalently applied to each reward, state, and action set resulting in a simplified exposition.

In the cyber security investment context, the model in equation (2.1) is analogous to the model by Parnell et al. (2015). An exception is that the probability distribution may depend on the time period, i, as for time-dependent formulations (Degroot (2005)). This time dependence persists throughout the observation, O, which is here assumed to be one of m levels, i.e., $O \in \{1, \ldots, m\}$. The key idea here is that the social media text is converted to a series of observations, $o_1, o_2 \ldots$, one for each period, with relevance to the decision problem. Then, the probabilities are updated using Bayes' theorem:

$$p_i(y, a_i|O = o) = \frac{p_0(y,a_i)p(O=o|y,a_i)}{\sum_{y=1}^{s} p_0(y,a_i)p(O=o|y,a_i)} \quad (2.2)$$

where p_0 is the initial or prior probability and the so-called "observation" matrix is $p(O = o|y, a_i)$ for indices o = 1, …, m, y =1, …, s, and each possible action a_i. Establishing the prior during the "burn in" Phase 1 is part of preparing for continuing fluctuations in Phase 2. The formulation in equations (2.1) and (2.2) is relevant for problems in which the system resets between periods, a phenomenon which applies only approximately to our cyber security case study.

The objective of the startup phase is to estimate the observation matrix, $p(O = o|y, a_i)$, using training data. Then, in steady state (Phase 2), the analysis method is used to generate observations, O, from the social media. Updates are performed using equation (2.2) and the result is used to solve equation (2.1) to generate the optimal action for the relevant time period *i*. In each period, action follows the observation.

2.2.2 Observations and Observation Matrices

The following sections describe a computationally efficient method to derive observations $O_1, \ldots, O_n$ over n periods for which the system states $y_1, \ldots, y_n$ are assumed known for known actions $a_1, \ldots, a_n$. Counts for the number of times an observation was observed in each state are $C_{o,a,y}$ for o = 1, …, m, a_i = 1, …, v, and y =1, …, s. Then, the observation matrix, $p(O|y, a_i)$ is estimated using

$$p(O = o|y, a) = \frac{C_{o,a,y}}{\sum_{o'=1}^{m} C_{o',a,y}}$$

$$\text{for } o = 1, \dots, m, a_i = 1, \dots, a, \text{and } y = 1, \dots, s \tag{2.3}$$

which derives the standard frequentist probability estimates. Observation matrices are displayed for each action, a_i, with rows corresponding to states, y, and columns corresponding to observation levels (Smallwood and Sondik (1973)). Observations are informative about the system state if the probabilities are concentrated along the columns of the observation matrices. Then, if the relevant observation level occurs, the Bayesian update in equation (2.2) generates a high probability that the system is in a specific state.

2.3 Efficient Methods for Obtaining Observations from Social Media

In this section, we review the LDA model which is a probability distribution from Blei et al. (2003). Then, we review the associated estimation methods from Blei et al. (2003) and Tey et al. (2006) and Griffiths and Steyvers (2004). In the next section, we propose a new estimation method based on transforming a k-means clustering model into an LDA model.

Note that virtually all text modeling methods begin with a natural language processing step in which text is transformed into numbers with irrelevant words removed and words "stemmed" (e.g., "jumping" and "jumps" both shorted to "jump" see Feldman and Sanger (2007) and Porter (1980)).

2.3.1 Latent Dirichlet Allocation

Our notation follows Blei et al. (2003) and Carpenter (2010) so that $w_{d,j}$ is the j^{th} word in d^{th} document with $d = 1, \ldots, D$ and $j = 1, \ldots, N_d$. Therefore, "D" is the number of documents or Tweets, and "N_d" is the number of words in the d^{th} document. We transform words into numbers using the method of Porter (1980). Therefore, $w_{d,j} \in \{1, \ldots, W\}$, where W is the number of distinct words in all documents.

The clusters or "topics" are defined by the estimated probabilities, $\hat{\phi}_{t,c}$, that a randomly selected word in cluster $t = 1, \ldots, T$ (on that topic) would achieve the specific value $c = 1, \ldots, W$. The value $\hat{\theta}_{d,t}$ represents the estimated probability a randomly selected word in document d is assigned to cluster t of the T possible. Estimating the $\hat{\phi}_{t,c}$ and $\hat{\theta}_{d,t}$ for $t = 1, \ldots, T$, $d = 1, \ldots, D$ and $c = 1, \ldots, W$ permits estimation of the observations needed for our timely decision analysis problem. This follows because we are interested in clusters or topics related to our problem by the probabilities, $\phi_{t,c}$ and periods in which the document probabilities, $\theta_{d,t}$, on these topics are high. The model variables $z_{d,j}$ are the cluster assignments for each word in each document, $d = 1, \ldots, D$ and $j = 1, \ldots, N_d$.

Generally, low values or diffuse prior parameters α and β are applied (Griffiths and Stuyvers (2004)). Note that these priors are relevant to Bayesian estimation of LDA only. The joint probability of the data, $w_{d,j}$, and the parameters to be estimated, $(z_{d,j}, \theta_{d,t}, \phi_{t,c})$, is (Carpenter (2010)):

$$P\left(w_{d,j}, z_{d,j}, \theta_{d,t}, \phi_{t,c} \middle| N_d, \alpha, \beta, d = 1, \dots, D, t = 1, \dots, T, c = 1, \dots, W\right) =$$

$$\left[\prod_{t=1}^{T} \frac{\Gamma\left(\sum_{c=1}^{W} \beta\right)}{\prod_{c=1}^{WC} \Gamma(\beta)} \prod_{c=1}^{W} \phi_{t,c}^{\beta-1}\right] \left[\prod_{d=1}^{D} \frac{\Gamma\left(\sum_{c=1}^{W} \alpha\right)}{\prod_{c=1}^{WC} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha-1}\right] \times$$

$$\left[\prod_{d=1}^{D} \prod_{t=1}^{T} \theta_{d,t}^{n_t^{(d)}}\right] \times \left[\prod_{t=1}^{T} \prod_{c=1}^{W} \phi_{t,c}^{n_t^{(c)}}\right]$$

where $\Gamma(\quad)$ is the gamma function and

$$n_t^{(d)} = \sum_{j=1}^{N_d} \sum_{c'=1}^{W} I\left(z_{d,j} = t \;\&c = c'\right) \quad \text{and}$$

$$n_t^{(c)} = \sum_{d=1}^{D} \sum_{j=1}^{N_d} I\left(z_{d,j} = t \;\&\; w_{d,j} = c\right) \tag{2.4}$$

and where $I(\dots)$ is an indicator function giving 1 if the equalities hold and zero otherwise. Note equation (2.4) is a simple representation of human speech in which words, $w_{d,j}$, are multinomial draws associated with given topics, $z_{d,j}$, which are also multinomial draws. The probabilities, $\phi_{t,c}$, that define the topics are also random, i.e., it is a hierarchical distribution. Technically, the estimates that are often used for these probabilities are Monte Carlo estimates for the posterior means of the Dirichlet distributed probabilities, $\hat{\phi}_{t,c}$.

Once the parameters $\hat{\phi}_{t,c}$ and $\hat{\theta}_{d,t}$ have been estimated, the derivation of the observations is relatively easy. Studying the estimated posterior mean probabilities of $\phi_{t,c}$, the clusters or topics (*t*) relevant to the decision problem are identified. Then, retrieving the documents on these topics with values of $\hat{\theta}_{d,t}$ that exceed a threshold in each time period, gives the needed observation counts, $O_1, \dots, O_n$. For example, if there are many Tweets on cyber vulnerabilities, the period is likely associated with elevated threats necessitating additional investment.

2.3.2 Collapsed Gibbs Sampling

Perhaps the most popular way to estimate the parameters in the LDA model in equation (2.4) is called "collapsed Gibbs" sampling (Teh et al. 2006, Griffiths and Steyvers 2004). To implement collapsed Gibbs, the values of the topic assignments for each word, $z_{d,j}$, are sampled uniformly. Then, iteratively, multinomial samples are drawn for each topic assignment $z_{d,j}$ iterating through each document, *d*, and word, *j*, using the last iterations of all other assignments, $z_{-(\boldsymbol{d},\boldsymbol{j})}$. The multinomial draw probabilities are

$$P\left(z_{d,j} = t \middle| d, j, z_{-(\boldsymbol{d},\boldsymbol{j})}, w_{d,j}\right) \propto \left(\frac{n_t^{(w_{d,j})} - I\left(z_{d,j}=t\right) + \beta}{n_t^{(\cdot)} - I\left(z_{d,j}=t\right) + W\beta}\right)\left(\frac{n_t^{(d)} - I\left(z_{d,j}=t\right) + \alpha}{n_\cdot^{(d)} - 1 + T\alpha}\right) \tag{2.5}$$

where $n_t^{(w_{d,j})} = \sum_{d'=1}^{D}\sum_{j'=1}^{N_d} I\left(z_{d',j'} = t \,\&\, w_{d',j'} = w_{d,j}\right)$,
$n_t^{(\cdot)} = \sum_{d'=1}^{D}\sum_{j'=1}^{N_d} I\left(z_{d,j'} = t\right)$,
$n_t^{(d)} = \sum_{j=1}^{N_d}\sum_{c'=1}^{W} I\left(z_{d,j} = t \,\&c = c'\right)$, and
$n_\cdot^{(d)} = \sum_{t'=1}^{T}\sum_{j'=1}^{N_d} I\left(z_{d,j'} = t\right)$.

In words, each word is randomly assigned to a cluster with probabilities proportional to the counts for that word being assigned multiplied by the counts for that document being assigned. After *M* iterations, the last set of topic assignments generate the estimated posterior means using:

$$\hat{\phi}_{t,c} = \frac{n_t^{(c)} + \beta}{n_t^{(\cdot)} + W\beta} \tag{2.6}$$

And the posterior mean topic definitions using

$$\hat{\theta}_{d,t} = \frac{n_t^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \tag{2.7}$$

Therefore, if words are assigned commonly to certain topics by the Gibbs sampling chain, their frequency increases the posterior probability estimates both in the topic definitions, $\hat{\phi}_{t,c}$, and the document probabilities $\hat{\theta}_{d,t}$. From $\hat{\theta}_{d,t}$ we can see periods when certain topics dominate.

2.4 K-means based Latent Dirichlet Allocation (KLDA)

Gibbs sampling is noisy and inefficient since only a single iteration of topic assignments is used for the posterior estimates and even approximate convergences can require thousands or millions of iterations. The proposed estimation method clusters documents. This is different from LDA, which permits documents to have specific words on multiple topics. Yet, for short documents like Tweets, the difference may be considered unimportant and robustness is explored in Section 2.6.

Denote the word counts for each document, d, and word, c, as $X_{d,c}$. The standard k-means clustering in our notation is (Lloyd (1982)):

1. Select T document, $d_1, \ldots, d_T$, uniformly from {1, …, D}, initialized the cluster centroids using $q_{t,c} = X_{d_t,c}$ for c = 1, …, W and t = 1, …, T.
2. Compute the distances for each document to each centroid using:

   $$v_{d,t} = \sqrt{\sum_{c=1}^{W}(q_{t,c} - X_{d,c})^2} \text{ for } t = 1, \ldots, T, d = 1, \ldots, D. \quad (2.8)$$

3. Assign each document to a cluster using, $\tilde{z}_d$, using

   $$\tilde{z}_d = \text{argmin}_t\, v_{d,t} \text{ for } d = 1, \ldots, D. \quad (2.9)$$

   S_t is the set of documents with $\tilde{z}_d = t$ for t = 1, …, T.

4. Update the centroids using the average locations for documents in the cluster:

$$q_{t,c} = \frac{\sum_{d \in S_t} x_{d,c}}{\sum_{d \in S_t} 1} \tag{2.10}$$

5. Repeat steps 2 through 4 until the cluster assignments do not change.

A last step is added to permit fractional membership in clusters by documents and facilitate the interpretation as a topic model. The "membership" function, similar way to fuzzy-c clustering, is:

$$u_{d,t} = 1/v_{d,t} \text{ for } t = 1, \ldots, T, d = 1, \ldots, D. \tag{2.11}$$

This permits estimation of the document topic probabilities using:

$$\hat{\theta}_{d,t} = \frac{u_{d,t}}{\sum_{d'=1}^{D} u_{d',t}} \text{ for } t = 1, \ldots, T, d = 1, \ldots, D \tag{2.12}$$

Also, the estimated topic definitions are generated using:

$$\hat{\phi}_{t,c} = \frac{q}{\sum_{c'=1}^{W} q_{t,c'}} \text{ for } t = 1, \ldots, T \text{ for } c = 1, \ldots, W \tag{2.13}$$

as the topic proportions, which show the distribution of topics in all the document lists. Clearly, if the documents are long and cover many substantially different topics, the approximation will be poor. We explore the robustness computationally in Section 2.6.

2.5 Cyber Security Twitter-Enabled Study

In this section, we use a routine decision problem faced by many organizations to illustrate the application of the formulation, modeling of social media data, observations, and results. The authors are aware of an organization that suffered losses perhaps exceeding $1M because of failure to solve this problem optimally. Often, organizations

do not attempt to patch medium-level cyber vulnerabilities. Patching requires staff time and can cause disruptions because some software may not work after patching actions.

Yet, during times of elevated risks resulting from exceptionally problematic medium-level vulnerabilities, adjustments are potentially relevant. Also, in these cases, the actions of administrators do not affect the threat level, but only the rewards (or losses). This simplifies our formulation in equation (2.1) since the probabilities do not depend on the actions. Twitter has experts tweeting continually on many subjects relevant to decision problems. The experts cover many topics and there are hundreds of potentially relevant medium-level vulnerabilities. Continued discussion of a medium vulnerability by the experts is likely an indicator of an elevated risk state.

Here, we study D = 16,047 tweets starting in January 2014 for 12 months from the 16 Twitter accounts relating to cyber security: Mathewjschwartz, Neilweinberg, Scotfinnie, Secureauth, Lennyzeltser, Dangoodin001, Dstrom, Securitywatch, Cyberwar, Jason_Healey, FireEye, Lancope, Varonis, DarkReading, RSAsecurity, and Mcafee_Labs. The decision problem includes s = 2 states (normal and elevated risk), a = 2 actions (1 – do not patch medium-level vulnerabilities, 2 – patch medium-level vulnerabilities). We assume that the system was in state 1 except for 4 months starting in April as indicated in Table 2 (a) because of the announcement of the well-known "Heartbleed" vulnerability. The database has W = 894 distinct words.

Applying k-means based LDA, topic 17 is identified as related to cyber vulnerabilities in general and "Heartbleed" in particular. It is the only topic for which one of the top 20 defining words is a medium vulnerability. The stemmed results for the top

words generated using equation (2.13) are shown in Table 1. Note how obscure our decision problem is with so much discussion being largely irrelevant and the need for filtering.

Then, KLDA identifies the top 20 documents by posterior mean estimate, $\hat{\theta}_{d,t}$, for each of the 12 months. Inspecting these Tweets manually and tabulating relevant mentions of "Heartbleed" (or any other medium vulnerability) resulted in the raw mentions in Table 2(a). In most periods, medium vulnerabilities received no mentions. For simplicity, observations are divided into two levels, i.e., level 1 – zero mentions of Heartbleed or level 2 – greater than zero mentions. This results in the observations $O, \dots, O_{12}$ and cross-tabulating generates the counts in Table 2(b) $C_{O,a,y}$. The frequentist estimates for the observation matrices are given in Table 2(c). The prior values and posterior estimates from equation (2.2) are provided for different observations in Table 2(d).

Table 1. The posterior mean topic definitions, $\hat{\phi}_{t,c}$, estimates from KLDA with 17 on Heartbleed

T1	0.0567	T2	0.0540	…	T17	0.0500	…
Word	Prob	Word	Prob	…	Word	Prob	…
(frequency) low	0.1393	(text) rt	0.1095	…	(name)	0.1040	…
(name) cyberwar	0.0291	(frequency) low	0.0927	…	(frequency) low	0.1018	…
(name)	0.0264	(name)	0.0183	…	(text) infosec	0.1009	…
(name) darkread	0.0182	(name) cyberwar	0.0169	…	(text)	0.0420	…
(month) 3	0.0164	(month) 2	0.0166	…	(frequency) medium	0.0192	…
(month) 2	0.0162	(name)	0.0165	…	(text) breach	0.0152	…
(month) 4	0.0154	(frequency) high	0.0132	…	(text) new	0.0152	…
(month) 1	0.0153	(name) jasonhealei	0.0124	…	(text) risk	0.0147	…
(name)	0.0135	(name) mcafeelab	0.0123	…	(text) malwar	0.0129	…
(month) 5	0.0131	(month) 3	0.0123	…	(month) 4	0.0125	…
(month) 8	0.0127	(text) secur	0.0112	…	(month) 5	0.0121	…
(month) 7	0.0120	(month) 1	0.0112	…	(text) attack	0.0121	…
(name) jasonhealei	0.0116	(text) atdavemarcu	0.0104	…	(text) hack	0.0116	…
(month) 6	0.0113	(month) 4	0.0100	…	(month) 6	0.0098	…
(month) 12	0.0099	(month) 8	0.0099	…	(text) secur	0.0098	…
(name) mcafeelab	0.0091	(month) 6	0.0094	…	**(text) heartble**	0.0098	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Table 2. (a) States $y_1, \ldots, y_{12}$, raw observations, and $O, \ldots, O_{12}$, (b) counts $C_{O,a,y}$, (c) observation matrices, $p(o|y,a)$, and (d) posterior values, $p(y,a|O)$, for different observation levels

(a)

Months	System State	Raw Mentions	Observation
1	2	0	1
2	2	0	1
3	2	0	1
4	1	7	2
5	1	4	2
6	1	1	2
7	1	0	1
8	2	0	1
9	2	1	2
10	2	0	1
11	2	0	1
12	2	0	1

(b)

state	1 (0)	2 (>0)
state 1	1	3
state 2	7	1

(c)

	1 (0)	2 (>0)
state 1	0.250	0.750
state 2	0.875	0.125

(d)

state	p_0	1 (0)	2 (>0)
state 1	0.333	0.125	0.750
state 2	0.667	0.875	0.250

We assume that attempting patching of mediums will reduce intrusions even while patching takes time and costs money. We assume rewards of $r(y = 1, a_i = 1) =$–\$300,000, $r(y = 2, a_i = 1) =$ \$100,000, $r(y = 1, a_i = 2) =$ –\$200,000, and $r(y = 2, a_i = 2) =$ –\$50,000 and piecewise utility $u(y) = 2y$ for $y < 0$ and $u(y) = y$ otherwise. If we observe $O_i = 1$ (no mentions of medium vulnerabilities), the expected utilities are $\rho(a_i = 1) =$ 0.125 and $\rho(a_i = 2) =$ – 1.375. With observation $O_i = 1$ (mentions of medium vulnerabilities), the expected utilities are $\rho(a_i = 1) =$ –4.250 and $\rho(a_i = 2) =$–3.245. Therefore, if the experts tweet about medium vulnerabilities, the

optimal action is patching. Otherwise, patching is not called for. This example illustrates how social media analytics can inform timely decision problems.

2.6 Numerical Studies

In this section, a computational comparison of Gibbs sampling and KLDA is provided. Four test corpora drawn from Allen et al. (2016) include two having multiple topics per document permitting the sensitivity of KLDA performance to be studied. The purpose to clarify the computational and accuracy advantages of the alternative estimation methods.

2.6.1 Test Problems

In this section, four similar cases are studied to compare different estimation methods. To preview, Table 2 summarizes the results of the computational runtimes. Table 6 (in the appendix) shows the four similar cases in which 40 documents are studied, so that $D = 40$ for each case. Table 7 (in the appendix) shows the true model topic proportion and topic definition, where topic number $T = 5$ for cases 1 and 2 and $T = 6$ for cases 3 and 4. The dictionary size for all the cases $W = 25$.

2.6.2 Evaluation Metrics

Because the estimated distribution topics have no natural ordering, it is hard to compare the result against the assumed ground truth. Therefore, Steyvers and Griffiths (2007) proposed that the permutations of cluster labels should be considered and the closest "distance" permutation should be selected. Define the function $t'(\mathbf{r}, t)$ as the selection of topic *t* in permutation **r**. Use $\phi_{t,c}^{true}$ to denote the ground truth topic definitions for *t* = 1, …, *T and* for *c* = 1, …, *W*. In the Appendix, the ground truth is provided for one of the four cases. For all cases, see Allen et al. (2016). Further, denote **r*** as the argmax permutation for equation (2.13). The accuracy measure used here is the average root mean squared (RMS):

$$RMS(\phi) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{c=1}^{W}\left(\phi_{t,c}^{true} - \phi_{t'(\mathbf{r}^*,t),c}\right)^2}. \tag{2.14}$$

Intuitively, the RMS value indicates the typical size of errors in the topic definition estimation.

2.6.3 Comparison Results

Figure 1 shows the comparison results of K-means LDA, Gibbs Sampling LDA with 10, 100, and 1000 runs. Each value in the table is the average RMS for 100 replications, i.e., starting from distinct random seeds. Using RMS metrics, K-means LDA could achieve a similar level of distance or even a smaller distance to the true model compared to other models. This holds even if there are multiple topics in each document (case 3 and case 4). For Gibbs sampling, Monte Carlo simulation introduces uncertainties.

A higher number of iterations gives slightly better RMS than lower numbers, but the quality is highly influenced by the random seed. Table 3 gives the timing for estimation methods. Clearly, KLDA is significantly more efficient with comparable quality. It permits our VBA software to analyze 10,000 Tweets in less than 20 minutes on an i5 processor.

Table 3. Computational accuracy (RMS) and timing results for the case studies

Case	Test Model	Iterations	Average RMS	Std RMS	100 Replicates Time (Sedc) (sec)
1	k-means LDA	2	0.0453	0.0000	5
1	Gibbs Sampling LDA	10	0.0507	0.0098	4
1	Gibbs Sampling LDA	100	0.0451	0.0089	44
1	Gibbs Sampling LDA	1000	0.0436	0.0064	323
2	k-means LDA	2	0.0500	0.0000	5
2	Gibbs Sampling LDA	10	0.0531	0.0076	6
2	Gibbs Sampling LDA	100	0.0492	0.0063	43
2	Gibbs Sampling LDA	1000	0.0492	0.0049	301
3	k-means LDA	2	0.0401	0.0000	6
3	Gibbs Sampling LDA	10	0.0482	0.0093	6
3	Gibbs Sampling LDA	100	0.0416	0.0063	56
3	Gibbs Sampling LDA	1000	0.0409	0.0046	489
4	k-means LDA	2	0.0450	0.0000	6
4	Gibbs Sampling LDA	10	0.0519	0.0080	7
4	Gibbs Sampling LDA	100	0.0456	0.0075	59
4	Gibbs Sampling LDA	1000	0.0459	0.0053	485

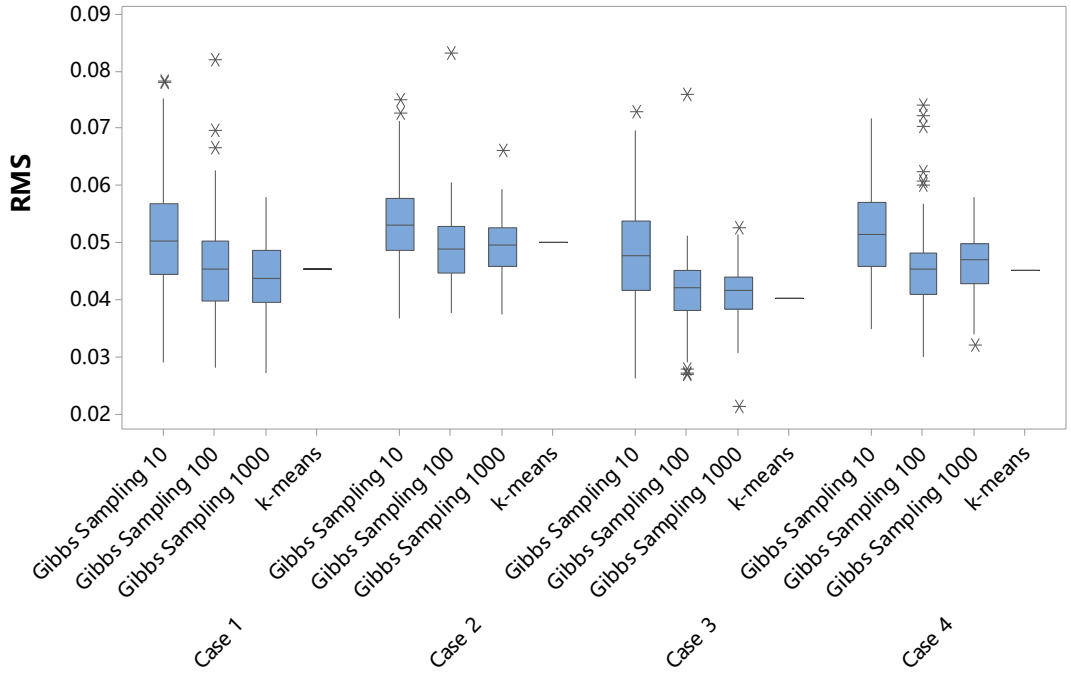

Figure 1. RMS comparison for different estimation methods for LDA only

2.7 Conclusions

In this chapter, we proposed a method to link social media analytics with routine decision analyses. We also proposed an innovative topic estimation technique based on k-means clustering called KLDA. This permits the rapid estimation of LDA models. The latter incorporate human high-level domain knowledge so that users can direct or perturb the model and results. Applying the techniques to test problems, we demonstrated that KLDA can achieve improved repeatability and comparable subjective accuracy. Specifically, we used four cases to test our new model against the true models. The

improved efficiency is important for enabling spreadsheet applications, allowing users to benefit from text processing and information retrieval for private text corpora.

2.8 References

Allen, T. T.; Sui, Z.; and Parker, N. (under 2nd review). "Timely Decision Analysis Enabled by Efficient Social Media Modeling". Submitted to *Decision Analysis*.

Allen, T. T.; Xiong, H.; and Afful-Dadzie A. (2016). "A Directed Topic Model Applied to Call Center Improvement". *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Blei, D.M.; Ng, A.Y.; and Jordan, M.I. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3, pp. 993-1022.

Bok, H.S.; Kankanhalli, A.; Raman, K.S.; and Sambamurthy, V. (2012). "Revisiting Media Choice: a Behavioral Decision-Making Perspective". *International Journal of e-Collaboration (IJeC)* 8(3), pp. 19-35.

Borrero, J.S.; Prokopyev, O.A.; and Saur é, D. (2015). "Sequential Shortest Path Interdiction with Incomplete Information". *Decision Analysis* 13(1), pp. 68-98.

Cao, Y. (2014). "Reducing Interval-Valued Decision Trees to Conventional Ones: Comments on Decision Trees with Single And Multiple Interval-Valued Objectives". *Decision Analysis* 11(3), pp. 204-212.

Carpenter, B. (2010). "Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling". revision 1.4, LingPipe,

Inc., carp@lingpipe.com, lingpipe.files.wordpress.com/2010/07/lda1.pdf (as of 5-4-2017).

Charalabidis, Y. and Loukis, E. (2012). "Participative Public Policy Making Through Multiple Social Media Platforms Utilization". *International Journal of Electronic Government Research (IJEGR)* 8(3), pp. 78-97.

DeGroot, M. H. (2005). *Optimal Statistical Decisions.* Vol. 82, John Wiley & Sons (reprint 1970).

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Gao, X.; Zhong, W.; and Mei, S. (2013). "Information Security Investment When Hackers Disseminate Knowledge". *Decision Analysis* 10(4), pp. 352-368.

Ghosh, S. and Dubey, S. K. (2013). "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms". *International Journal of Advanced Computer Science and Applications* 4(4).

Griffiths, T. L. and Steyvers, M. (2004). "Finding Scientific Topics". *Proceedings of the National Academy of Sciences* 101(suppl 1), pp. 5228-5235.

Lee, S. H. (2012). *Comparison and Application of Probabilistic Clustering Methods for System Improvement Prioritization*. Doctoral dissertation, The Ohio State University.

Lloyd, S. P. (1982). "Least Squares Quantization in PCM". *IEEE Transactions on Information Theory* Vol. 28, pp. 129–137

Madsen, R. E.; Kauchak, D.; and Elkan, C. (2005). “Modeling Word Burstiness Using the Dirichlet Distribution”. In *Proceedings of the 22nd International Conference on Machine Learning*. pp. 545-552.

Miller, S.; Wagner, C.; Aickelin, U.; and Garibaldi, J. M. (2016). “Modelling Cyber-Security Experts' Decision Making Processes Using Aggregation Operators”. *Computers & Security* 62, pp. 229-245.

Packiam, R. M. and Prakash, V. S. J. (2015). “An Empirical Study on Text Analytics in Big Data”. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).* pp. 1-4.

Parnell, G. S.; Butler III, R. E.; Wichmann, S. J.; Tedeschi, M,; and Merritt, D. (2015). “Air Force Cyberspace Investment Analysis”. *Decision Analysis* 12(2), pp. 81-95.

Pat é-Cornell, M. E. (2012). “Games, Risks, and Analytics: Several Illustrative Cases Involving National Security and Management Situations”. *Decision Analysis* 9(2), pp. 186-203.

Porter, M.F. (1980). “An Algorithm for Suffix Stripping”. *Program* 14(3), pp. 130-137.

Smallwood, R. D. and Sondik, E. J. (1973). “The Optimal Control of Partially Observable Markov Processes over a Finite Horizon”. *Operations Research* 21(5), pp. 1071-1088.

Steyvers, M. and Griffiths, T. (2007). “Probabilistic Topic Models”. *Handbook of Latent Semantic Analysis* 427(7), pp. 424-440.

Sui, Z.; Milam, D.; Allen, T. T. (2015). “A Visual Summarizing Technique Based on Importance Score and Twitter Feeds”. *INFORMS Social Media Analytics Student Paper Competition*.

Sun, X. (2014). “Textual Document Clustering Using Topic Models”. In *Semantics, Knowledge and Grids (SKG), 2014 10th International Conference on*. pp. 1-4.

Teh, Y.W.; Newman, D.; and Welling, M. (2006). “A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation”. In *Advances in Neural Information Processing Systems*, pp. 1353-1360.

Von Neumann, J. and Morgenstern, O. (2007). *Theory Of Games And Economic Behavior*. Princeton University Press (2nd ed. 1947).

Zhao, T.; Li, C.; Li, M.; Wang, S.; Ding, Q.; Li, L. (2012). “Predicting Best Responder in Community Question Answering Using Topic Model Method”. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology* 01, pp. 457-461.

CHAPTER 3
SUBJECT MATTER EXPERT REFINED TOPIC (SMERT) MODEL AND A VISUAL SUMMARIZING TECHNIQUE BASED ON IMPORTANCE SCORE AND TWITTER FEEDS

3.1 Introduction

Zaman, Herbrich, Gael, and Stern (2010) described some of the many possible uses of Twitter and other social media. For example, companies and research institutes are using social media to predict how events will be received based on the thoughts and feelings users are posting. Some Hollywood production companies are using Twitter to predict how a movie will perform. Its use may have also helped improve how particular movies have done in the box office (Britt (2015)). Here, we explore the use of Twitter-based analysis methods for improving sensemaking and summarizing. The specific case study examples that we use relate to improving the situation-awareness of system administrators in cyber security contexts. Cyber-security is a growing field of study due to the growing use of data collection and the use of newer internet enabled devices. Therefore, this paper will investigate through examples of the connection between cyber-security and social media, in particular Twitter, in addition to their individual importance.

Yang and Counts (2010) studied Twitter and used name association and key word identification to track the speed with which tweets travel through accounts and the paths

that these tweets take. Their study finds that mentioning an individual on Twitter indicates that a tweet will have more diffusion in terms of speed and number of users viewing and reposting said tweet. Zaman, Herbrich, Gael, and Stern (2010) presented a method for predicting the spread of information in a social network using retweets as positive feedback and lack of retweets as negative feedback. The number of retweets can be used as an important indicator in the prediction model for social events and changes. Zaman, Fox, and Bradlow (2014) used a Bayesian approach to develop a probabilistic model for the evolution of retweet counts. Their model successfully predicted the final total number of retweets through the time-series path of retweets. In our examples, we use retweet counts as an indicator of importance and our point-based "importance score" can be viewed as an approximate estimate of retweet counts.

Building on Latent Dirichlet Allocation (LDA), Allen, Xiong, and Afful-Dadzie (2015) proposed subject matter expert refined topic (SMERT) for probabilistic clustering of texts to permit experts or users to edit the topics using knowledge about the system or their own needs. SMERT and LDA estimate the proportion of words in the overall corpus on each topic. As a special case of LDA, SMERT potentially incorporates "high-level" inputs from a subject matter expert to adjust the topics and clusters by zapping or boosting words in the topic definitions. Allen, Vinson, Raqab, and Allam (2013) applied the SMERT model to course evaluation analysis. Using Pareto charts, this method helped to screen out less effective feedback and allow researchers to focus on relevant and important information.

Topic models and SMERT have shown promise for creating intuitive summaries of bodies of text. But there are issues with estimation and in particular topic proportions are often poorly estimated and fail to capture what is new temporarily in the topic proportions. Therefore, this article proposes a visualization and point-base system designed to help users with sense-making of Twitter feeds. The goals of this paper are to overcome the reported estimation issues from the SMERT models and demonstrate the value of the point system in relation to Twitter summarizing. To illustrate the problems with SMERT and LDA and the advantage of the proposed TWitter Importance Score Topic (TWIST) modeling, we seek to show improved correlation with retweet counts of the point system. In Section 3.2, we describe a motivating example relating to cyber vulnerabilities and describe the need for interpretation. In Section 3.3, we will review SMERT models. In Section 3.4, we propose the point system associated a visualization method, which could aid in many Twitter-related sense-making cases. In Section 3.5, we return to the cyber-security case studies and illustrate the application of the proposed methods. Finally, we summarize our findings and suggest opportunities for future research.

3.2 Cyber Vulnerability Example

To demonstrate the new TWIST method, we use a case from 2014 relating to cyber security. During 2014, there were several major cyber vulnerabilities that became public knowledge. Most notably was the vulnerability commonly known as the Heartbleed. The Heartbleed vulnerability was made public knowledge on April 1, 2014.

This vulnerability resulted from a lack of bounds in memory allocations for operating systems. The vulnerability and notification allowed for large amounts of information to be stolen from any susceptible computer. Upon this disclosure many hackers made use of the vulnerability before a patch could be created. As a result the number of attacks on a large Midwest institution's computers increased by approximately 400% in the month of April as shown in Figure 2.

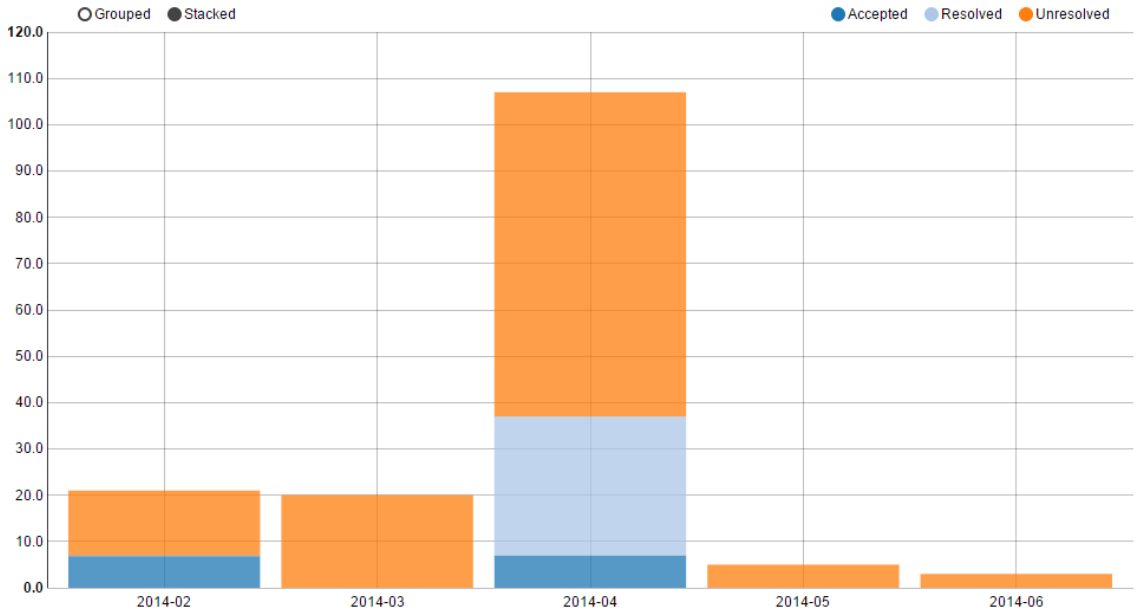

Figure 2. Known computer intrusions for a large Midwest organization in 2014

No doubt, some system administrators at the Midwest organization knew about Heartbleed after the announcement but many did not. Yet, all observed the spike in attacks as detected using the intrusion detection system (IDS). The IDS generally intercepts only a fraction of all attacks so likely some were missed and all administrators needed to perceive the vulnerability and understand its cause. This is the objective of our

proposed methods in this article, i.e., to improve situation awareness at all times by synthesizing Twitter feeds into an intuitive chart.

As another example, we consider the November 2014 attack on the Sony Corporation by, reportedly, North Korea. This was a well-publicized attack that received a large amount of media attention. These famous cyber events raise discussions on social media platforms. The proposed methods seek to identify and interpret the events in both cases.

3.3 Subject Matter Expert Refined Topic

Allen, Xiong, and Afful-Dadzie (2015) proposed subject matter expert refined topic (SMERT) for probabilistic clustering of texts. Both are "topic" models with the topics being clusters of words in the documents associated with fitted multivariate statistical distributions. In practice, not all of the distribution is relevant to the user and the topics can be represented by ordered lists of words which users often find interpretable. SMERT is a generalization of Latent Dirichlet Allocation (LDA) methods.

SMERT generalizes LDA in that it incorporates input from a Subject Matter Experts (SMEs) or ordinary users. The method derives the main topics with a body of documents and estimate what portion of the text corresponds to each topic. Extended from equation 2.4 of LDA, SMERT has a distribution as equation (3.1). The distribution is fit using collapsed Gibbs sampling which is a form of Markov Chain Monte Carlo. Collapsed Gibbs is an iterative process where the topic assignments and distribution are modified. The topic assignments converge to the samples from the new distribution and

are then used for estimations for the topics and proportions. Below you will find equation 3.1, or how fitting the distribution.

$$P\left(w_{d,j}, z_{d,j}, \theta_{d,t}, \phi_{t,c} \middle| N_d, \alpha, \beta, d = 1, \dots, D, t = 1, \dots, T, c = 1, \dots, W\right) =$$
$$\left[\prod_{t=1}^{T} \frac{\Gamma\left(\sum_{c=1}^{W} \beta\right)}{\prod_{c=1}^{WC} \Gamma(\beta)} \prod_{c=1}^{W} \phi_{t,c}^{\beta-1}\right] \times \left[\prod_{d=1}^{D} \frac{\Gamma\left(\sum_{c=1}^{W} \alpha\right)}{\prod_{c=1}^{WC} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha-1}\right] \times$$
$$\left[\prod_{d=1}^{D} \prod_{t=1}^{T} \theta_{d,t}^{n_t^{(d)}}\right] \times \left[\prod_{t=1}^{T} \prod_{c=1}^{W} \phi_{t,c}^{n_t^{(c)}}\right] \times$$
$$\left[\prod_{t=1}^{T} \prod_{c=1}^{W} \binom{N_{t,c}}{x_{t,c}} \phi_{t,c}^{x_{t,c}} \left(1 - \phi_{t,c}\right)^{N_{t,c} - x_{t,c}}\right] \quad (3.1)$$

Based on equation 2.4, what is new is that $x_{t,c}$ and $N_{t,c}$ are collected from a boost and zap table, and $x_{t,c}$ is the successes out of $N_{t,c}$ Bernoulli trials for all topics $t = 1, \dots, T$ and words $c = 1, \dots, W$. They are the high-level input form SMEs. Then the expert's domain of knowledge could accurately identify the topics and topic definitions from the boost and zap table of the output results of initial inspection of LDA results.

- For boosting, $N_{t,c} \geq x_{t,c} > 0$. In this case, the experts' domain of knowledge decides the word has a nonzero probability in the topic, which affirms the word would appear in the topic. For example, in the research, $N_{t,c} = x_{t,c} = 2$ can be treated as a strong affirmation of the word being in the topic with 2 out of 2 draws from the topic.
- For zapping, $N_{t,c} \geq x_{t,c} = 0$. In this case, the experts' domain of knowledge decides the word has a zero probability in the topic for most of the cases, which affirm the word do not belong to the topic. For example, in the

research, $N_{t,c} = 100{,}000$ and $x_{t,c} = 0$ can be treated as a strong belief of the word being in the topic with 0 out of 100,000 draws from the topic.

The left two rectangles in the graphical model representation of Figure 3 shows the conditional relationships between the variables in the LDA model. In the figure, the rectangles indicate the number of elements in each random vector or matrix. For example, the matrices *z* and *w* have N_d elements for each of the *D* documents.

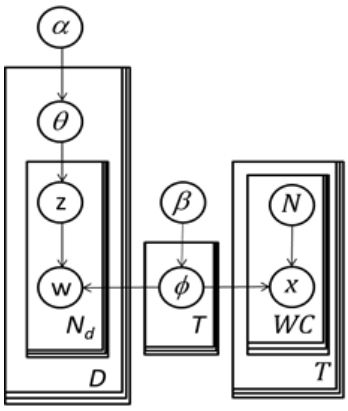

Figure 3. Graphical model of SMERT and LDA. LDA is the left two most portions

The posterior mean values for the topic definitions ϕ and topic proportions θ are estimated through a single replicate of the topic assignments after some level of convergence (Blei, Ng, and Jordan (2003)). These posterior means give a conceptual map of the corpus because the words with highest probabilities in the topic definitions offer the most meaningful cluster definitions. The proportions for each topic summarize the corpus and prioritize later visualization parts (Steyvers and Griffiths (2007)).

3.3.1 K-means based SMERT Model

Followed Section 2.4, a K-means based SMERT is proposed for the topic model under supervision more efficiently.

$$\phi'_{t,c'} = \frac{\sum_{d=1}^{D} N_d \times \phi_t \times \phi_{t,c} + x_{t,c'}}{\sum_{d=1}^{D} N_d \times \phi_t + x_{t,c}}, \quad t = 1, \dots, T, c' = 1, \dots, W \tag{3.2}$$

$$\phi'_t = \frac{\sum_{d=1}^{D} N_d \times \phi_t + x_{t,c'}}{\sum_{d=1}^{D} N_d + x_{t,c}}, \quad t = 1, \dots, T, c' = 1, \dots, W \tag{3.3}$$

$$\theta_{d,t}{}' = \frac{N_d \times \theta_{d,t} + \frac{n_d^{(c)}(d = d' \& c = c')}{\sum_{d=1}^{D} n_d^{(c)}(d = d' \& c = c')} \times x_{t,c'}}{N_d + \frac{n_d^{(c)}(d = d' \& c = c')}{\sum_{d=1}^{D} n_d^{(c)}(d = d' \& c = c')} \times x_{t,c'}}$$

$$d = 1, \dots, D, t = 1, \dots, T, c' = 1, \dots, W \tag{3.4}$$

To replicate SMERT with K-means methods, I propose the topic definition and topic proportion as SMERT with equation 3.2 – 3.4. To explain the equations in words, equation 3.2 replicate the topic definitions. In the boost table, we choose to boost the word by $x_{t,c'}$. Using the output of topic definition and topic proportion, the expected number of word c' in topic t is calculated by $\sum_{d=1}^{D} N_d \times \phi_t \times \phi_{t,c}$. So all the words are added this topic and then the expected number becomes $\sum_{d=1}^{D} N_d \times \phi_t \times \phi_{t,c} + x_{t,c'}$. Then the updated topic definition $\phi'_{t,c'}$ is calculated by equation 3.2 and the update topic proportion ϕ'_t is calculated by equation 3.3. To zap the word, $\phi'_{t,c'}$ is arbitrarily assigned to be zero.

For sentence-topic probability matrix, $n_d^{(c)}(d = d' \& c = c')$ gives out the number of word c' in sentence d'. Then sum it up to get the total number of word c' in the corpus. Then $\frac{n_d^{(c)}(d=d' \& c=c')}{\sum_{d=1}^{D} n_d^{(c)}(d=d' \& c=c')} \times x_{t,c'}$ calculate the number of word c' that is added to sentence d'. Then update the sentence-topic probability by $\theta_{d,t}{}'$ of equation 3.4 To zap the word, $\theta_{d,t}{}'$ is arbitrarily assigned to be zero.

3.4 Twitter Importance Score Topic Summarizing Method

3.4.1 Notations and Assumptions

In this section, we define additional notations and assumptions for the proposed TWIST method. Consider a finite number of text document with *D* sentences and each sentence is signified as d, where $d = 1, \ldots, \mathrm{D}$.

Our method is based on the SMERT method but it could be based on LDA only. In either case, the derived topics are denoted t_i, $\forall i \in I$, where *I* is the set of topic indices. Within each topic, the words are ordered as w_{ij}, $\forall j \in \mathrm{J}$ and J is the set of word indices. P_{il} is the estimated mean posterior probability (what the Gibb sampling generates) that sentence l falls in the topic t_i. A set of documents is called a corpus and q is the number of top words in each topic that are studied by the subject matter expert. The default is q = 10 words for each topic (top 10). Also, the predicted score or importance number is the *PS*.

Our method is based on the SMERT method but it could be based on LDA only. In either case, the derived topics are denoted t. Within each topic, the words are ordered as c. $\theta_{d,t}$ is the estimated mean posterior probability or the sentence-topic probability matrix (what the Gibb sampling generates) that sentence d falls in the topic t. A set of documents is called a corpus and q is the number of top words in each topic that are studied by the subject matter expert. The default is q = 10 words for each topic (top 10). Also, the predicted score or importance number is the *PS*.

3.4.2 The Proposed Twitter Importance Score Topic (TWIST) Summarizing Method

The TWitter Importance Score Topic Summarizing Method is as follows.

---

0. Select Twitter content to follow and create a corpus of tweets from the relevant time period.
1. Run LDA on the corpus.
2. Loop over each topict
    a. Loop over each word $c \in the\ first\ q\ words\ in\ the\ topic$, zap c if c does not make sense. otherwise, boost it. End loops.
3. Run SMERT without sorting topics using the high-level boosts and zaps.
4. Loop over each tweet (sentence) d with property m, $V_{tm} = \sum_{l \in m} \theta_{d,t}$
5. Loop over each topic $m \in M$, rank V_{tm} from largest to smallest
6. Select *N* largest values of V_{tm}, $V_{nm} = V_{tm}$, where $n = 1, \dots, N$
7. For all $m \in M$ and $n = 1, \dots, N$,
    a. If count of topic t, $Count_t = 1$, assign predicted score $PS_{1tm} = Count_{1n}$, where $n = 1 \dots N$, $Count_{11} > Count_{12} > \cdots > Count_{1N}$.
    b. If count of topic t, $Count_t = 2$, assign predicted score $PS_{2tm} = Count_{2n}$, where $n = 1 \dots N$, $Count_{21} > Count_{22} > \cdots > Count_{2N}$.
    c. If count of topic t, $C_i \neq 1$ or 2, assign predicted score $PS_{tm} = 0$.
8. $S_m = \sum_t (PS_{1tm} + PS_{2tm} + PS_{tm} + PS)$, where PS is the constant predicted score for all $m \in M$. End loops.
9. Plot PS_{1tm}, PS_{2tm}, PS_{tm}, PS in a column chart with short phrase extracts from the topic definitions as labels.

TWitter Importance Score Topic (TWIST) Summarizing Method

For Step 4, $V_{tm} = \sum_{d \in m} \theta_{d,t}$ means sum all of the probabilities for the same property. Here, the property includes examples like different months, years, or even days. For SMERT, normally 20 topics are selected as outputs. In Step 6, among the 20 topics, normally, $N = 5$, or the top 5 topics are selected in the TWIST summarizing method in most cases. In the method, predicted scores are normally either equal to predicted numbers or proportional to the predicted numbers.

3.5 Case Studies

In this section, two cyber Tweet examples are studied using the TWIST summarizing method. The first case study relates to the Heartbleed vulnerability from 2014. The second relates to the Sony Hack and, to a lesser extent, the Shellshock vulnerability which occurred in the same time period. In this section, these examples are presented to show how the TWIST summarizing method could react to new topics. Admittedly, both case studies could have been studied together but we wanted to evaluate the level of generality of TWIST and explaining them separately is simpler.

The same 15 Twitter broadcasters were analyzed for the purposes of both studies (*Step 0*). These users were found by searching for the Twitter users who have a reputation for being cyber-security analysts. Also, a combination of individuals and organizations/groups were found to ensure there wasn't a bias based on the goal of the Twitter user. The Twitter sources (usernames) in the following examples are: Mathewjschwartz, Neilweinberg, Scotfinnie, Secureauth, Lennyzeltser, Dangoodin001,

Dstrom, Securitywatch, Cyberwar, Jason_Healey, FireEye, Lancope, Varonis, DarkReading, RSAsecurity, and Mcafee_Labs.

3.5.1 Heartbleed

3.5.1.1 Background

Heartbleed is security vulnerability in the windows software package. As discussed earlier, the vulnerability was made public knowledge in April 2014 with wide ranging impacts. Many attacks were attempted before a patch could be applied to remove the vulnerability. The Heartbleed received a large amount of publicity due to the severity and number of people it impacted (millions).

3.5.1.2 Data and Computation Results

Figure 4 shows the total number of retweets for the first six months of 2014 which will be compared later to the chart this new method generates. Notice that the retweet counts correlate with the known intrusion counts confirming that retweet counts often relate to important events.

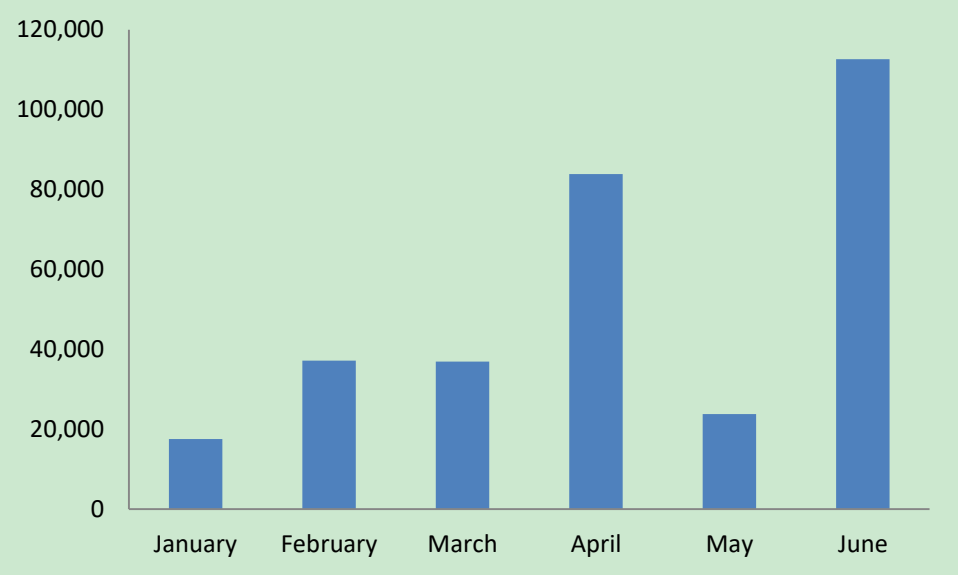

Figure 4. Retweets for January to June 2014 with the Heartbleed announcement in April

Next, we applied the remaining steps in the TWIST method. Below are the topics that SMERT created based upon the tweets and zapping any unwanted words. Steps 1-3 involve applying SMERT. We zapped Heartbleed in February and March because we know from our expertise that there were no tweets about Heartbleed until April after it was publically announced. The developed topics were then identified by words and the words. Then, we manually translated the word lists into (hopefully) interpretable topics with the results in Table 4.

Table 4. SMERT topics which were interpreted manually incorporating the highest frequency words

Number	Topics
1	Jason Healey and cyberwar, among others, tweeted with a moderate following tweeted in a few months about cyber security.
2	RSA security, among others, tweeted about its own products and an event called the archer summit with a small following.
3	Dangoodin001, among others, retweeted topics from many different months, without much following on Twitter.
4	Cyberwar, among others, tweeted about Eric Snowden and the NSA in multiple months
5	MacAfee Lab, Darkread, and dstrom, among others, tweeted about network security it multiple months
6	Dangoodin001 and Darkread, among others, tweeted about the Heartbleed with a moderate following on Twitter in particular during April.
7	Security watch and dangoodin001, among others, tweeted about apps and passwords with a moderate following on Twitter.
8	Lancop tweeted about its own company in particular during February and March with a low number of retweets.
9	Mathewjschwartz and Darkread, among others, tweeted about the target breach and information security with a low number of retweets.
10	Lennyzeltser and security watch, among others, retweeted topics and specifically at a neiljrubenk on Twitter.
11	Mathewjshwartz and dangoodin001, among others, tweeted with a high number of retweets in multiple months.
12	RSA security and Darkread, among others, tweeted about data security in multiple months
13	Fire eye, among others, tweeted about information security, malware, and threats with a moderate number of retweets particularly in April and May.
14	Varoni, among others, tweeted about information security and data privacy in multiple months.
15	Varoni, among others, tweeted about big data and security in multiple months with a low number of retweets.
16	Scotfinnie and security watch among others tweeted about Microsoft windows with a low number of retweets
17	Cyberwar and Dangoodin001 among others tweeted about thanking others in multiple months
18	Varoni and Darkread, among others, tweeted about social media and information security in multiple months.
19	McAfee Lab, among others, tweeted about security stories and particularly to Twitter users davemarcu and Slashdot in multiple months with a low number of retweets.
20	Darkread and Varoni, among others, tweeted about information security and Darkread in particular during April and June.

For both examples we use N = 5. Also, for steps 4-8, the predicted score is $PS = 10{,}000$. If the topics are unique among all the 6 month, a predicted score of 65,000, 52,000, 39,000, 26,000, and 13,000 is assigned if the topic ranks No. 1 to 5 respectively. If the topics appear twice among all the 6 month, assign predicted score of 15,000, 12,000, 9,000, 6,000, and 3,000 if the topic ranks No. 1 to 5 respectively. Figure 5 shows the predicted scores with a breakdown by topics (Step 9).

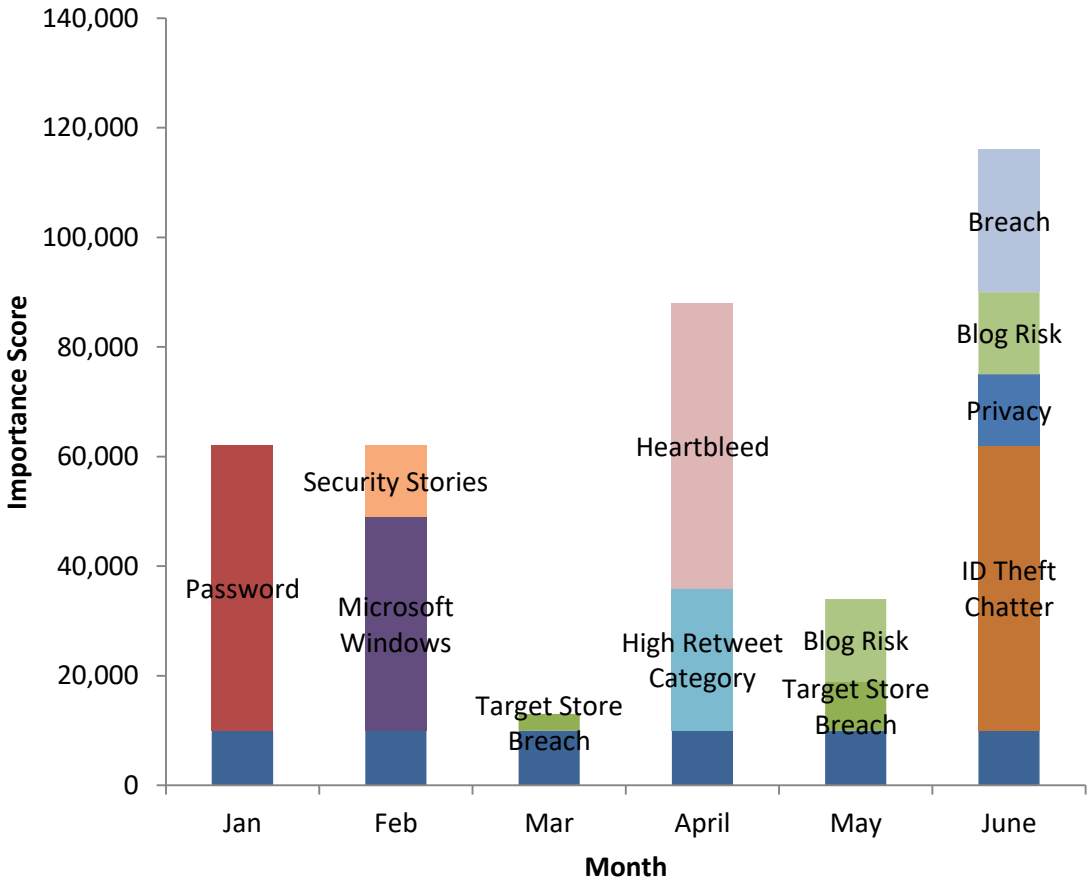

Figure 5. Heartbleed example predicted score breakdown by topic

3.5.1.3 Discussion

As can be seen in the figure above, April is characterized by the Heartbleed event and the high retweet category. This makes sense, as the Heartbleed event would cause a few particular announcements and updates to be highly retweeted. This characterizes April as a month which is abnormal and focuses on the Heartbleed event.

January and February both have slightly elevated PS and predicted retweet numbers as well. The January password focus and February story and system update focus may result in part from the Target credit card theft in December of the previous year, and an increased focus on cyber security. The target theft involved many peoples credit card information being stolen and was a major event for many individuals who may not think of cyber security very often.

The month of June also had a large number of points associated with it. June seemed to have a large amount of discussion associated with breaches of security resulting in theft of personal information and privacy issues. However, this system did a good job of predicting the real retweet numbers. The month of April was clearly dominated by discussions of the Heartbleed vulnerability which is exactly what an IT professional would want to know about if they did not know already.

3.5.2 Shellshock and the Sony Hack

3.5.2.1 Background

Shellshock is a security bug in Unix Bash Shell. It was disclosed on September 24, 2014. Many web server deployments use Bash to process web requests. Therefore, the bug could cause potential vulnerability issues to execute arbitrary commands and allow attackers acquiring unauthorized access to hosts. This bug can be compared to the Heartbleed bug in severity as it could potentially compromise millions of unpatched hosts.

The Sony Hack is another interesting example as it aroused more public attention. However, it is probably less relevant to local system administrators. On November 24, 2014, Sony released a movie called *The Interview*, which is about North Korea and their leader's dictatorship. Therefore, North Korea attacked Sony's online system and hacked Sony employees' personal data. Both of these two events attract active discussions on social media.

3.5.2.2 Data and Computation Results

The data set is also from the 16 Twitter accounts as in the Heartbleed example, but the data in these two examples are from July to December in 2014.

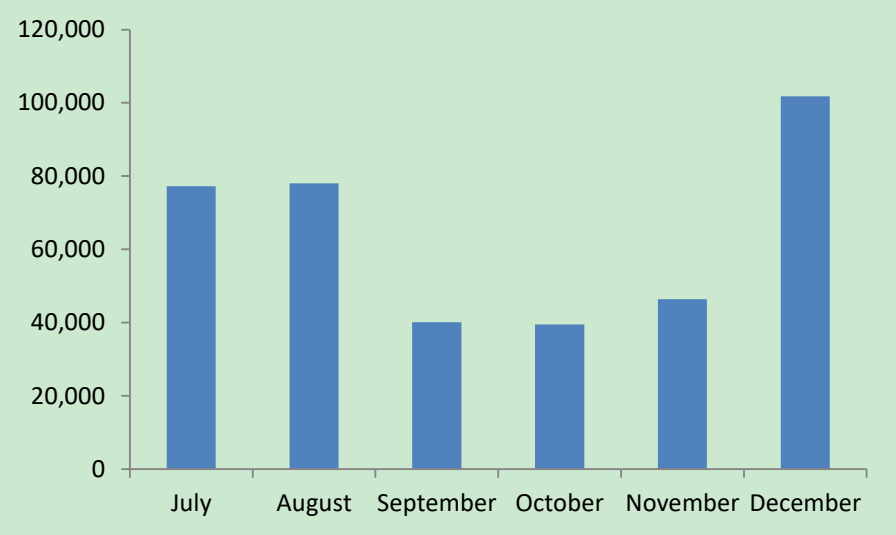

Figure 6. Retweet numbers from the period involving Shellshock and the Sony hack

Figure 6 shows the total retweet number across the 16 accounts. Different from expectations about Shellshock and the Sony Hack events, the total retweet numbers in September and November are not very high compared to other months.

Table 5. SMERT topics for the second case study during the Shellshock and Sony hack period

Number	Topics
1	Lancop tweeted about information security and cyber security for companies during July, August, October, and November with high number of retweets.
2	Lennyzelts, varoni and nealweinberg tweeted about new malware tool in October and December with high number of retweets.
3	Varoni tweeted about big data, information security, and data privacy in July and August with high number of retweets.
4	Mathewjschwartz, darkread and scotfinni tweeted about malware breach for Apple during September to November with high number of retweets.
5	Dangoodin001 and lennyzelts tweeted about year 2014 in August and December with high number of retweets.
6	Cyberwar and jasonhealei tweeted and retweeted about new things on internet during July, August and November with high number of retweets
7	Mathewjschwartz, cyberwar and dangoodin001 tweeted and retweeted about the Sony Hack during December with high number of retweets.
8	Dstrom, mathewjschwartz and cyberwar tweeted and retweeted about great reading and look during September and November with high number of retweets.
9	Secureauth tweeted and retweeted about security authenticity during September and October with high number of retweets.
10	Securitywatch tweeted and retweeted about online ID security protection during October and November with high number of retweets.
11	Jasonhealei tweeted and retweeted about cyber attack and National Security Agency (NSA) during September and October with high number of retweets.
12	Securitywatch, dangoodin001 and mathewjschwartz tweeted and retweeted about apps on mobile device during July, August and November with high number of retweets.
13	Fireeye tweeted and retweeted about information security during July, August and October with high number of retweets.
14	Dangoodin001 tweeted about thank and questions during July, November and December with high number of retweets.
15	Darkread and dstrom tweeted about cloud data breach and security during July, October, and November with high number of retweets.
16	Rsasecur tweeted about blog, sharing security and RSA summit event during September and December with high number of retweets.

Continued

Table 5 continued

17	Cyberwar, darkread, and mathewjschwartz tweeted about the new bug Shellshock and potential attack during August to October with high number of retweets.
18	Rsasecur tweeted about cyber security threat detection in RSA during October and November with high number of retweets.
19	Mcafeelab tweeted about malware attack and new phishing threat report during July and December with high number of retweets.
20	Varoni tweeted about information security and password hack during July and August with high number of retweets.

After zapping the unwanted words, SMERT output 20 topics as in Table 5. Topics 7 is about the Sony Hack and Topic 17 is about Shellshock. In this example, the parameters and importance scores are assigned as the same values from the previous example. Then using the TWIST summarizing method, Figure 7 shows the predicted scores with a breakdown by topic for the Shellshock and the Sony Hack example.

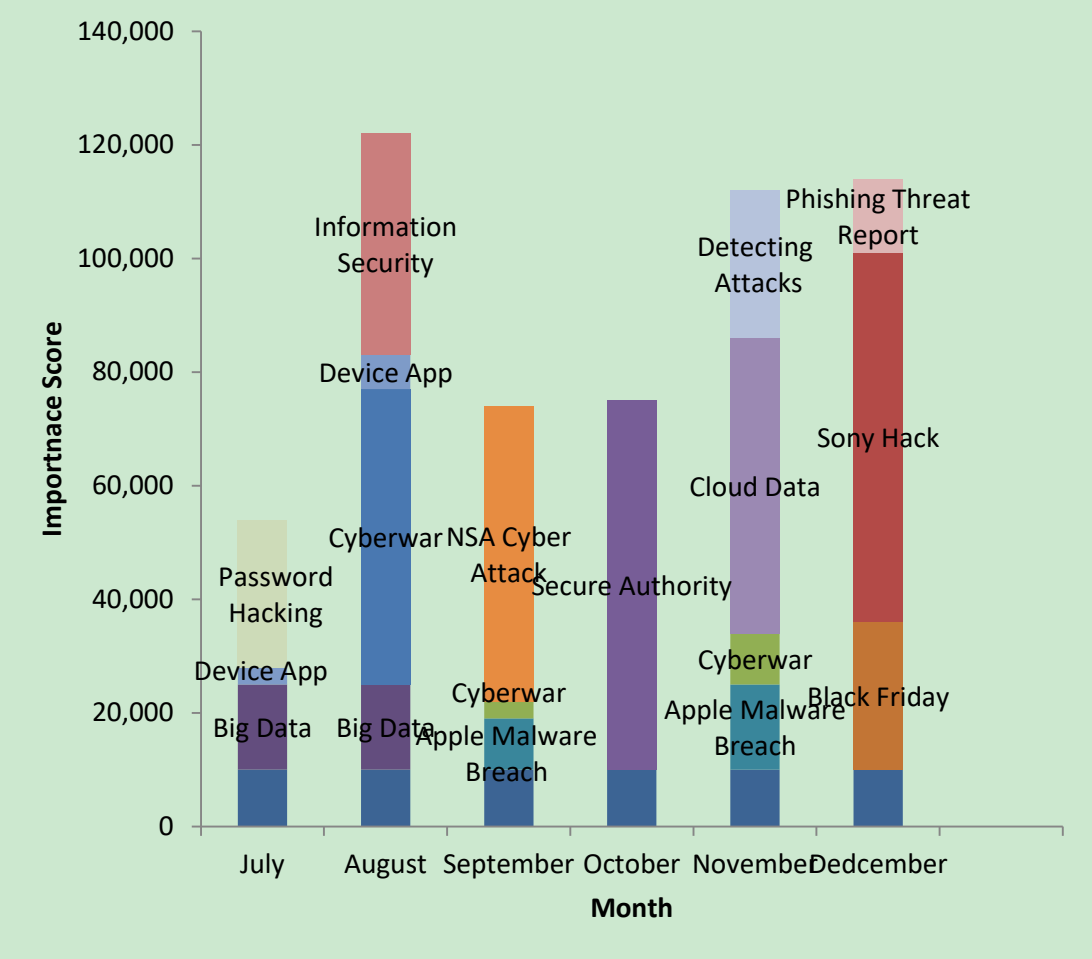

Figure 7. Shellshock and the Sony Hack Predicted Scores Breakdown by Topics

3.5.2.3 Discussion

In Figure 6, July, August and December have a higher total retweet number than other months. The reason that July and August have a higher total retweet number may be from discussion of the Unix Bash Shell (shell shock) on social media. Shellshock did not receive its name until September however. The reason that December has a higher total retweet number may be because the Sony Hack happened in late November. Although it

aroused active discussions on social media in November, the total retweet does not react to this accident very sensitively due to the late time of the month. But the total retweet number of December behaves as we would expect.

Figure 7 shows that the predicted scores from the TWIST summarizing method are more sensitive to the social events than the real total retweet number. There is a peak in the August predicted score and the breakdown of topics for August has shown that the social media users have observed a new information security issue. As discussed in the last paragraph, the bug was just not named as Shellshock yet. The Shellshock bug being referred to consistently over the time frame means it does not show up as clearly using this method however.

3.6 Conclusions

In this article, we proposed the TWitter Importance Score Topic (TWIST) method to aid in summarizing and sensemaking. We illustrated the application of TWIST to two data sets related to cyber security. In the first case, the TWIST method explained at a glance the large uptick of cyber intrusions during the month of April 2014. The chart clearly shows that the uptick corresponded to the Heartbleed vulnerability. Similarly, for second case study, the Shellshock vulnerability is also readily apparent. Another relevant occurrence (the Sony Hack) is clearly visible. In both case studies, the so-called "importance score" correlated highly with the numbers of retweets providing confirmation that the TWIST method generates relevant information.

3.7 References

Allen, T.T.; Xiong, H.; and Afful-Dadzie A. (2016). "A directed topic model applied to call center improvement". *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Allen, T. T.; Vinson, S. M.; Raqab, A.; and Alam, Y. (2013). "Using SMERT to Identify Actionable Topics in Student Feedback." *Integrated Systems Engineering Technical Report 2013*.

Britt, R. (2015). "How you and 'The Rock' Turned His Movie Around". Retrieved from http://www.marketwatch.com/story/how-hollywood-is-using-social-media-to-tell-if-a-movie-will-be-a-hit-Accessed June 19, 2015.

Shah, D. and Zaman, T. (2010). "Community Detection in Networks: the Leader-Follower Algorithm". *arXiv preprint arXiv:1011.0774*.

Sui, Z.; Milam, D.; and Allen T.T. (2015). "A visual summarizing technique based on importance score and Twitter feeds". *INFORMS Social Media Analytics Student Paper Competition. Finalist.*

Yang, J. and Counts, S. (2010). "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter". *ICWSM* 10, pp. 355-358.

Zaman, T.; Fox, E. B.; and Bradlow, E. T. (2014). "A Bayesian Approach for Predicting the Popularity of Tweets". *The Annals of Applied Statistics* 8, pp. 1583-1611.

Zaman, T. R.; Herbrich, R.; Van Gael, J.; and Stern, D. (2010). "Predicting Information Spreading in Twitter". In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS* 104, pp. 17599-601.

CHAPTER 4
EXPLORATORY TEXT DATA ANALYSIS FOR QUALITY HYPOTHESIS GENERATION

4.1 Introduction

In this chapter, we propose Exploratory Text Data Analysis (ETDA) as a technique for use in the analysis of text-based data sets to generate and test hypotheses relating to system improvement. Quality engineers often have text data available on multiple subjects. These could be in the form of customer surveys, complaints, line transcripts, maintenance squawks, or warranty reports. Yet, they often lack the techniques to use this data effectively for quality improvement purposes.

Tukey (1977) proposed Exploratory Data Analysis (EDA) as a general method for generating hypotheses using visualizations for statistical problems. De Mast and Trip (2007) proposed a prescriptive framework for applying EDA in the context of quality improvement projects. Here, we focus on EDA in the context of both quality improvement and text data. Therefore, ETDA is intended to be a special case of EDA and the associated quality framework of De Mast and Trip (2007). Beyond providing just a set of techniques or data visualization methods, ETDA like EDA seeks to provide a coherent philosophy about how to perform data analysis (Tukey, 1977).

As noted by Tukey (1977) and others, EDA contrasts with Confirmatory Data Analysis (CDA). EDA seeks to generate hypotheses while CDA has the goal of testing existing hypotheses. For instance, in a regression/hypothesis testing problem, EDA might be conducted as a first step to identify possible regressors to include in a model using scatter or XY plots. The shot size in injection molding, e.g., might be hypothesized to affect the fraction on nonconforming units. Once the model form is selected, then CDA proceeds to calculation of the p-values and interpretation of their implications for proving hypotheses. Then, proof might be generated that shot size does indeed affect the fraction of nonconforming units.

Another type of analysis called Descriptive Data Analysis (DDA) is potentially used as part of both EDA and CDA (De Mast and Trip (2007)). DDA is concerned with the summary of data, e.g., statistics such as the sample mean and sample standard deviation. DDA also suppresses the uninformative part of the set to highlight its important features. In large-scale problems dealing with big data arrays, measurements such as means and standard deviations, visualizations in tables and graphs, or other descriptive statistics reduce the complexity of the data sets (Good (1983)). DDA helps inquirers to prune unimportant data and focus on the salient features. In the context of our proposed ETDA framework, preprocessing of data may be viewed as DDA. Therefore, like EDA, ETDA is intended to be an extension of DDA.

As noted previously, Exploratory Text Data Analysis (ETDA) is proposed to be a special case of EDA that analyzes plain text datasets to derive high-quality information in quality improvement topics. Allen and Xiong (2012) and Sui (2017) provide examples of

the application of ETDA techniques in exploring Toyota Camry user reviews. Allen et al. (2016) leverage ETDA for hypothesis generation for call center improvement factors. Here, we seek to provide a prescriptive framework for ETDA for all quality improvement projects using real-life applications of ETDA from case studies like those used by De Mast and Trip (2007). These cases are selected to represent a variety of areas relevant to quality practitioners including automotive engineering, calling centers, and information technology.

The following section proposes the ETDA framework and describes its relationship to the framework from De Mast and Trip (2007). Subsequent sections elaborate on the steps of the ETDA: preprocessing of the text data: text data analysis and display options; text data salient feature identification; and lastly salient identification interpretation. Finally, we offer remarks relating to the discussion of the issues that a practitioner might encounter while employing ETDA.

4.2 The Principles and Framework of ETDA

In this section, we review the purposes of EDA from De Mast and Trip (2007) and describe the special context of text modeling and ETDA. As noted by those authors, EDA's main purposes are "to generate hypotheses", "to generate clues", "to discover influence factors", and "to build understanding of the nature of the problem." Like the EDA framework, the ETDA framework also seeks to reveal the potential relationships between key process output variables (KPOVs) in six sigma terminology or Y's and the

associated key process input variables (KPIVs) or X's. Thus, the first principle formalizes the purpose:

A. *The purpose of ETDA is to leverage text documents to help in the identification of dependent variables, Ys, and independent variables, Xs, that may prove to be of interest for understating or solving the problem under study.*

De Mast and Trip (2007) draw a distinction between situations in which there is a relatively easy way to identify the KIVs (Ys) and other situations. When it is easier to differentiate (situation #1), the total negative instances, e.g., defects, are the sum of available categories of instances:

$$\mathrm{Y} = Y_1 + Y_2 + Y_3 + \cdots \tag{4.1}$$

In these situations, EDA (and ETDA) should be able to identify the leading terms and associate hypotheses for clear follow-up activities. The first example below shows how ETDA identifies dependent variables (Ys) for further study using natural language processing (NLP) and a popular clustering method called Latent Dirichlet Allocation (LDA, Blei, Ng, and Jordan, (2003)). Additional details about NLP and LDA are described in the next section.

Example 1: Quality Improvement for Toyota Camry

In the first example, the Toyota Camry consumer report dataset from Allen and Xiong (2012) is used. This dataset contains 1,067 records of user reports for the automobile model between the years 2000 to 2010. The data resemble customer

complaint or survey results available in many industries and was provided by Consumer Reports.

The records include fields of summary, pros, cons, comments, driving experience and so on. Here, only cons texts are analyzed to generate quality hypotheses. NLP and LDA are applied using 10 topics. Again, NLP and LDA are discussed in more detail in the next section and the appendix. The resulting, Pareto chart is given in Figure 8. In the chart, the clusters or topics are represented by the top words ranked using estimated posterior probability. The charted quantities are the estimated posterior cluster proportions following Allen and Xiong (2002).

The first topic can be interpreted to mean that consumers are complaining about road noise or wind noise because of tire problems. This topic accounts for 21.80% of con words among the ten topics. The second most frequent topic is about uncomfortable seating, accounting for 17.85%. This is consistent with the 2010 Camry recalls for seat heater/cooler problems caused by damage to electrical wiring in the seat heater when the seat cushion is compressed. The third most frequent topic (yellow column) verifies the well-known uncontrolled acceleration problem which embarrassed the Toyota Corporation during the 2009-2011 period.

The implications for quality improvement projects are clear. The data and charting aid in clarifying the appropriate priority level for addressing the widely publicized unintended acceleration problem over tire noise and uncomfortable seats. Even complete remediation of the unintended acceleration would reduce only approximately 17% of the claims.

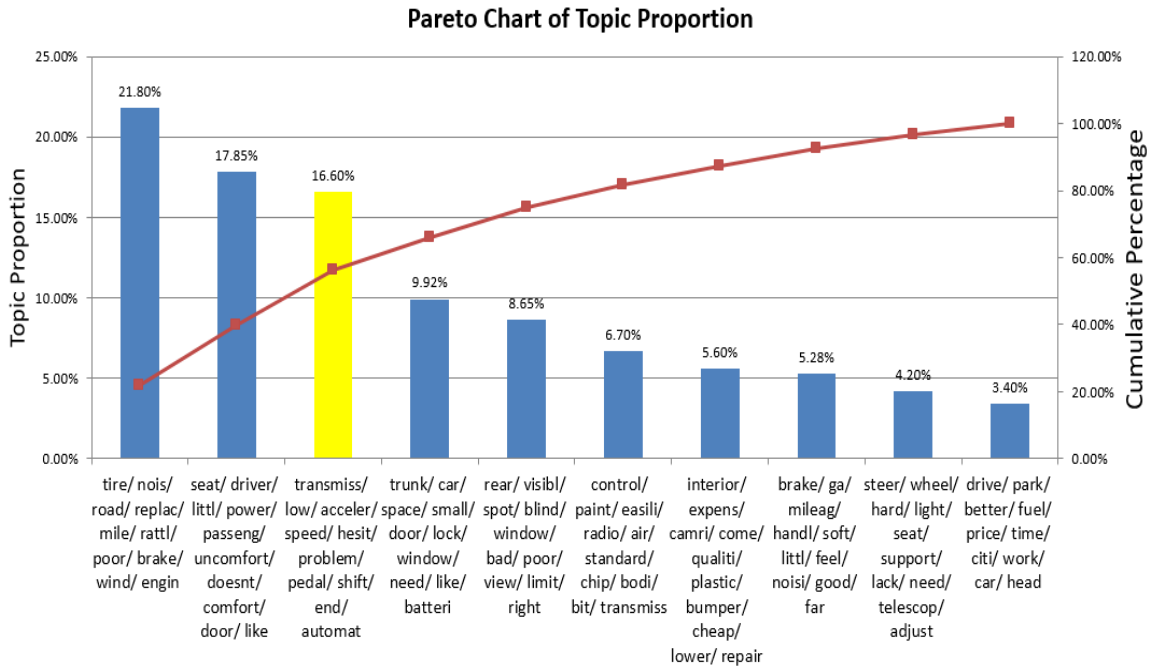

Figure 8. Topic Proportion for Cons in Toyota Camry Consumer Report

In the second type of situation classified by De Mast and Trip (2007), the data is more limited. The practitioner can only identify that there is another lower level of attribution and analysis needed. Then, the sum of negative events is written:

$$\mathrm{Y} = E_1 + E_2 + E_3 + \cdots \tag{4.2}$$

where $E_i = f(X_{j1}, X_{j2}, X_{j3}, \ldots)$. The investigators could acquire clues about causal factors (E_i) by analyzing the text documents with respect to independent variables (Xs). A case study of how ETDA helps to identify clues for further investigation about independent variables (Xs) is presented later in Example 2.

De Mast and Trip (2007) proposed a three-step process for quality improvement-related EDA. Because of the complexities of text modeling, in ETDA the first step of their process is divided into two parts creating four steps:

1. Text data preprocessing.

2. Text data analysis and display.

3. Salient feature identification.

4. Salient feature interpretation.

The next sections describe additional principles elaborating on those in De Mast and Trip (2007) for these steps.

4.3 Text Data Preprocessing

Natural language processes (NLP) methods are becoming increasingly standard (Feldman and Sanger (2007)). There are variants, of course. Yet, in general, irrelevant or "stop" words are removed such as "of" and "a" which often offer limited contributions to meaning. Then, words are "stemmed" so that "qualities" and "quality" might become "quality" and, potentially, synonyms are replaced. Finally, the stemmed words are replaced by numbers for clustering or other analysis activities. While NLP is an entire field of inquiry with many possible complications, the general emphases of Tukey and EDA are transparency and simplicity (Tukey (1977)). Therefore, the second ETDA principle is as follows.

B1. *NLP methods for ETDA should be simple with stop words that can be adjusted and standard stemming. Then, the users should perceive NLP as transparent and understandable.*

After the stop words are removed and the words are stemmed, a list of distinct words is called a "dictionary" for each set of documents. A simple approach used here to

address multiple fields in a database is to append field titles to these stemmed non-stop words. Porter (1980) proposed an algorithm to handle words that have different forms for grammatical reasons as well as derivationally related words with similar meanings. Combining all these steps the methods used in the example are:

Step 1. Split the document into words.

Step 2. Remove the punctuation or symbols and (optionally) make all words lower case.

Step 3. Remove the stopping words.

Step 4. Stem the words with the Porter Stemming Algorithm.

Step 5. Append the field titles in parentheses to each word (if appropriate).

Once the dictionary is available and the words are pre-processed, clustering and assignments of weights or "semantic" analysis are generally the primary techniques for additional processing. These methods create word or document cluster "tags" and numerical values to permit further data exploration steps. This leads to the principle:

B2. *Apply clustering methods to tag documents with numbers relating to cluster membership. These tags are useful for plotting and hypothesis generation.*

Among the most widely cited and used methods for unstructured text clustering is Latent Dirichlet Allocation (LDA, Blei et al. (2003)). LDA is described in more detail in the appendix. LDA involves fitting a distribution to the words with random probabilities (probabilities on probabilities) describing the chances that a random word is in a cluster (or "topic") and that it will assume a specific selection from the dictionary. Of primary interest are the probabilities defining the clusters or topics ("topic probabilities") and the

probabilities relating to the changes that words in specific documents relate to the topics ("document-topic" probabilities). The estimated mean values for these defining probabilities provide inputs to further analyses.

The direct Bayesian approach to estimate these mean posterior probabilities defining the clusters is called "collapsed Gibbs" sampling (Griffiths and Steyvers (2004); Teh et al. (2006)). Allen et al. (2017) created an approximate but relatively computationally efficient method for estimating the topic probabilities and the document-topic matrices based on k-means clustering. In this method, the clusters are used as topics by calculating the Euclidean distance from each quantified document to the estimated cluster centers and using the inverse of distance as the probabilities of the stemmed words falling in each topic.

Some clusters might be associated with problems or customer complaints of specific nature, as we illustrate in Figure 8. Yet, in general, words in topic models do not have clear positive or negative interpretations. In many situations, methods that explicitly place values on words in the dictionary can facilitate additional insights. This permits the study of quality issues at a higher granularity than the cluster level.

B3. *Apply a simple and relatively transparent sentiment score analysis to transform the text to values (positive, zero, or negative numbers for further analysis.*

There are many methods for assigning values to individual words, sentences, or documents relating to their positive or negative value (Pang and Lee (2008); Liu (2012); Turney (2002)). Generally, words are related to emotional states such as "angry", "anxious", "happy", "sad", or neutral. With arbitrariness, words can be rated individually

with scores about their strengths. Here, to reduce the arbitrariness and for simplicity, words are generically rated as positive or negative. The sum of the positive words in a document is denoted by *P* and the sum of the negative words by *N*. The sentiment score (*S*) used in our examples is

$$S = \ln(0.5 + P) - \ln(0.5 + N) \tag{4.3}$$

In general, our objectives for clustering and for sentiment scoring are to produce quantitative data to facilitate hypothesis generation. In the next section, we describe how the derived outputs can be used to create visualizations to aid in quality improvements.

4.4 Text Data Analysis and Display

After preparing text-creating numbers relating to cluster identities and membership or sentiment score, one can follow steps 2-4 which derive from methods of De Mast and Trip (2007). Then, graphical presentations in ETDA can aid in highlighting and presenting findings to analysts (Good (1983), Hoaglin et al. (1983), Bisgaard (1996)). Therefore, after preprocessing, the next step is to display text data in a straightforward way that exploits the power of pattern recognition.

C. *Process and display the quantitative text data to reveal distributions and potential hypotheses for ways to improve system quality*.

At this phase, the primary visualization tools include Pareto or sorted bar charting methods, running charts and so on to view different topics, contents, and quantitative text data.

C1 (Stratified Data). *Process and display the quantitative text data so as to reveal distribution across and within strata.*

Quality ratings can provide strata. Displaying cluster information at different strata levels can generate hypothesis for design inputs as illustrated in Example 2.

Example 2: Quality Improvement for Honda Civic

This example relates to 628 records for Honda Civic model between the years 2000 and 2010. The data are associated with the scores from 1 (very dissatisfied) to 5 (very satisfied). The field of comments and rating scores are used for data display and analysis. To improve customers' satisfaction, analysts are assigned to look for the reasons for low ratings from customers. Gibbs sampling estimation of LDA modeling are employed to cluster the comments into 10 topics.

Next, line charts of topic proportions for the documents associated with the different rating scores are presented in Figure 9. Only five top cluster definitions are plotted as rated by their proportions. The green line topic has 25.35%, 20.51%, 22.01%, 10.07%, and 5.50% for ratings from 1 (the poorest) to 5 (the best) respectively. This topic relates to complaints that "Honda has a transmission problem and needs to be repaired or replaced." Inspecting Figure 9, the hypothesis that focusing on transmission warrantee production would likely remove the major negative causes at all levels. It also suggests that a variety of levels of distress may be attributable to the same transmission cause.

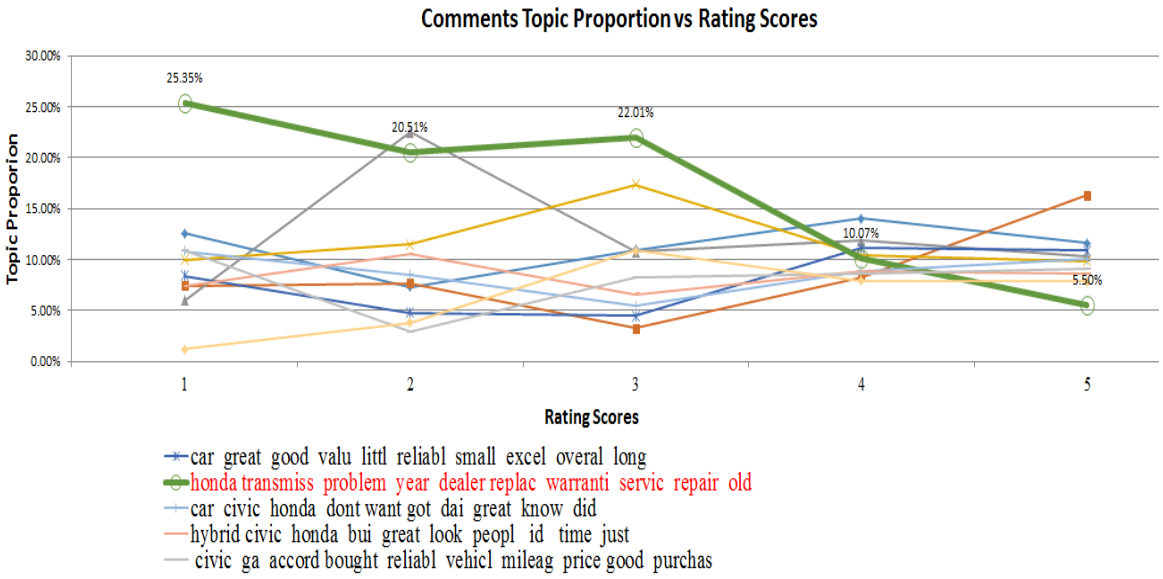

Figure 9. Topic Proportion vs Rating Scores for Comments in Honda Civic Consumer Report

Therefore, the decision variables associated with transmission design (Xs) are targeted for prioritization in design changes. The strata (rating scores) for different topics are the "variable containers" (De Mast and Trip (2007)), and the causal relationship is suggested through the variations of topic proportions across strata.

A special type of strata explored by De Mast and Trip (2007) is time strata. From their analysis, the following principle us derived:

C2 (Data plus time order). *Process and display the quantitative text data such that they will reveal distribution throughout the whole-time duration.*

This principle is illustrated in the following example. The example also illustrates roles for regression modeling, histogram, and trend plotting.

Example 3: Quality Improvement for Ford F-150

This example is also based on Consumer Reports data providing 369 records for the Ford F-150. The example involves the years 2000 to 2010 and the actual numbers of recalls from 2000 to 2010 during those years. Sentiment analysis is done for each of the 369 comments in the consumer report using equation (3) tabulated using software from CX Data Science. The linear relationship of sentiment scores and the actual number of recalls is shown in Figure 10(a). From this, the following linear regression model is derived:

$$\hat{S} = 0.4950 - 0.03767\ (\#\text{Recalls}) \tag{4.4}$$

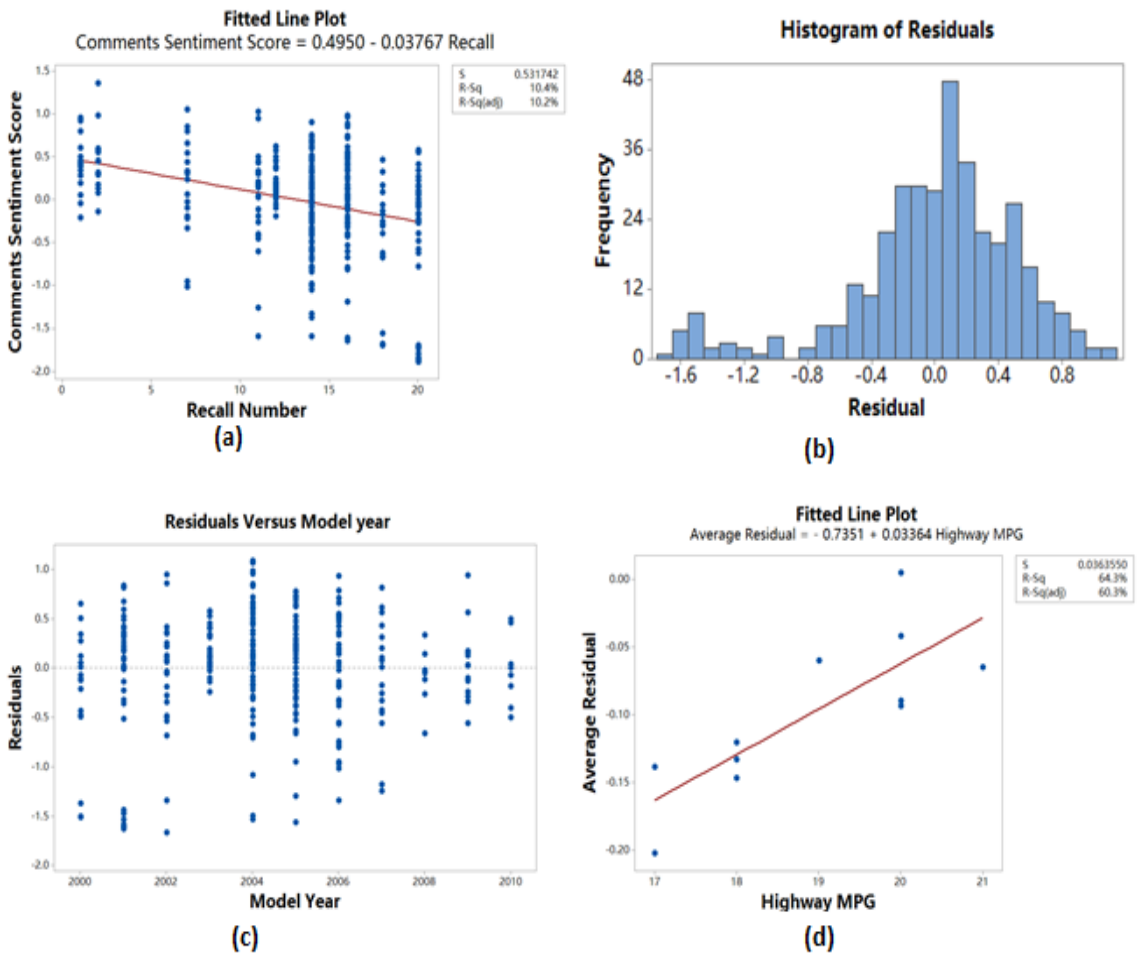

Figure 10. Linear Regression Model and Residual Plots for Comments Sentiment Scores

The sentiment score is predicted to be higher when the recall number is high. The residual plot of the linear regression relationship is plotted in the Figure 10(b). Based on the bimodality of the distribution for the residual plot, it seems that there is likely another cause for the low scores in addition to recalls.

Plotting the residuals by time strata (model year) in Figure 10(c) provides information about the timeliness of the missing cause. Most importantly, perhaps, the residual plot indicates that the causes do not endure to the latest model years.

The highest negative sentiment score residuals are found in 2000 and 2002. Exploration of the comments in 2000 and 2001 shows that many are about poor gas mileage. In 2000, the trucks miles-per-gallon averaged only 13 city/17 highway miles per gallon (MPG). From 2002 to 2007, truck MPG improved resulting in fewer customer complaints about this shortcoming. This is reflected in the less negative residuals in 2003-2007. After 2008, fuel consumption improved further, reaching more than 20 MPG on highways. The Figure 10(d) shows the linear relationship between yearly average residuals of the comments sentiment score versus MPG. Combining both the inferences from Figure 10(b) and Figure 10(d), it is suggested that further improvements might not reduce negative sentiment after 2010 since mean negative sentiment is dominated by recalls.

Example 3 illustrates how ETDA can provide insights relevant to design teams and related prioritization. This is a case in which text data is used to help discover independent variables (Xs) by focusing on one or more time intervals in the data. The bimodality distribution of residuals deviates from the expected normal distribution of linear regression residuals.

C3 (Multiple field data). *Process and display the quantitative text data to reveal distributions for different fields.*

In relation to the Toyota Camry case explored in Example 1, Allen and Xiong (2012) presented topic modeling across multiple data fields including summary, pros, cons, comment, and driving experience. To handle the multiple field data, the words in the dictionary are labelled with the field labels, e.g., “(summary) wear” which increases

the size of the dictionary but does not affect the mechanics of clustering in the appendix. An alternative way to handle multiple field text data would be to plot the causal relationships for all the fields on the same chart and compare them to look for variations within or across fields.

4.5 Text Data Salient Feature Identification

The next step in EDTA is the identification of the salient features again following EDA in De Mast and Trip (2007). Those authors wrote that salient features are the "finger prints" that clarify the key Xs and causes. Shewhart (1931 and 1939) defined the identification of salient features as finding out "the clues to the existence of assignable causes" for the non-randomness. The causes being sought, therefore, often relate to deviations of system outputs from standards or predicted outputs. This leads to the following principle:

D. *Search for deviations or variations from reference standard.*

Text data are different from normal numerical data in that they typically do not conform to certain distributions and contain a good deal of noise information. However, if certain causes of variation dominate, they would still leave clues for their identification. Also, the scales used such as sentiment analyses contain arbitrariness. As an example of this principle consider the residual analysis in Example 3 in Figure 10(b). The deviation signals another cause.

Another type of variation is between groups. This leads to the following principle.

D1 (Stratified data). *Look for deviations or variations from other groups.*

In the Honda Civic's consumer report in Example 2 it is seen that while the green line topic has a decreasing trend from rating score 1 to 5, most of the other topics have either a flat or an increasing trend. In Example 2 clearly, the green line differs greatly from other groups. This deviation of trending behavior reveals clues of salient features for the quality problem. This leads to clarity about the importance of transmission issues over other "groups" or types.

Another type of deviation relates to time periods leading to the following principle.

D2 (Data plus time order). *Look for deviations or variations from previous time intervals.*

Time series plots of cluster posterior probabilities (proportions) or sentiments can facilitate the search for important inputs (Xs). These could include partial autocorrelation function or, alternatively, simple difference plots. Example 4 illustrates the use of a run chart of the period-to-period differences in the posterior mean topic proportions revealing a salient feature. In this case, the salient figure relates to a new cause generating cyber security incidents.

Example 4: Cyber Attack Pike due to Heartbleed

This example uses the cyber security Twitter account data detailed in Allen and Xiong (2012) and Sui et al. (2015). A large Midwest institution suffered from a high number of cyber attacks and experienced a sudden computer intrusion hike in April 2014. To leverage ETDA for the quality hypothesis generation, inquirers collected 16,047

Tweets from January 2014 to December 2014 from 16 Twitter accounts of noted cyber experts. Gibbs Sampling Topic Modeling techniques is used to break the Tweets into 10 topics. For each topic, the topic proportions are acquired for each month and the differences from the previous month are charted in the running time chart in Figure 11. The third topic in the legend is associated with the grey-colored line and references the famous “Heartbleed” vulnerability.

Cluster or topic 3 experiences a sudden increase in topic proportion in April 2014 and a sudden decrease in topic proportion in October 2014, while other topics’ changes are relatively constant, fluctuating around zero. This pattern is consistent with the timing of the public disclosure of the vulnerability, “Heartbleed”, on April 1, 2014. This vulnerability resulted from a lack of bounds in memory allocations for operating systems, which allowed large amounts of information to be stolen from any susceptible computer.

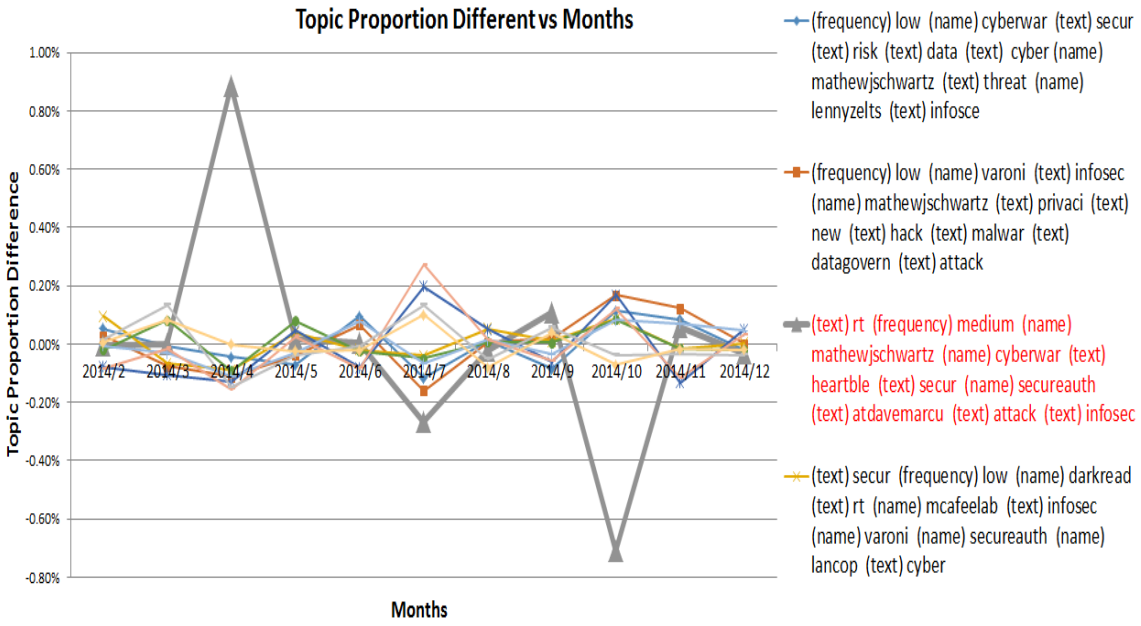

Figure 11. Topic Proportion Difference vs Months

Upon this disclosure, many hackers made use of the vulnerability before a patch could be created resulting, among other disruptions, in the roughly 400% increase in cyberattacks experienced by the large Midwest institution in the month of April 2014. The sudden increase in topic proportion that month shows a surge in discussion of the issue on Twitter. Figure 11 suggests both that the uptick in incidents might likely have been caused by Heartbleed and that the issues was resolved by September.

Example 4 also shows how cluster posterior probability estimates can provide reference values for comparisons between clusters. General reference comparisons lead to the principle:

D3. *Look for deviations or variations from other references which could be other fields.*

For data with multiple fields, a comparison of causal relationships across and within fields can be beneficial. If one or some of the fields behave differently from most other fields, the salient features of those abnormal fields could be explored for quality hypothesis generation.

D4. *Look for salient feature based on prior perceptions, rules, or knowledge.*

Prior experience or field knowledge can help to identify salient features. In Example 2 above, the finding that a topic has a high proportion of low ratings and a low proportion of high ratings, obviously suggests it is worth exploring for salient features related to quality.

4.6 Text Data Salient Feature Interpretation

In this step, salient features are turned into hypotheses using context knowledge. Niiniluoto (1999) introduced the concept of abduction in which "the inquirers compare conceptual combinations to the observations until all the pieces seem to fit together and a possible explanation pops up." As described in De Mast and Trip (2007), the final principle of the framework is:

E. *The identified salient features should be interpreted using context knowledge.*

Example 5 illustrates the use of context knowledge to generate hypotheses in a customer support feedback call center. In this case, the context knowledge enters explicitly in the clustering process. Also, it enters in identifying the subsystems that should likely be prioritized for additional study.

Example 5: Call Center Improvement

Allen et al. (2016) presented a call center service improvement problem to an insurance company. Using 2,378 records of conversations between the service representatives and callers, Allen et al. (2016) applied an extended topic modeling of subject matter refined topic (SMERT) to acquire 10 topics. SMERT generalizes topic modeling incorporating inputs from Subject Matter Experts (SMEs) by allowing analysts to decide whether certain words belong to the topics or not, and then to boost or zap words in the topic based on their domain knowledge. This is a method of refining techniques to filter information through pruning, generalizing or suppressing approaches to achieve discovery optimization.

Allen et al (2016) describes the binomial thought experimental data used to generate the refined model using approximate Gibbs sampling. It also details the use of expert knowledge to re-label the clusters. Instead of merely using the words associated with the highest posterior mean probability estimates, sentences are used for labeling. This permits immediate generation of hypotheses. For example, studying Figure 12, we estimate that improved automatic information and verification (relating to topic 2) might reduce the call volume by approximately 12%.

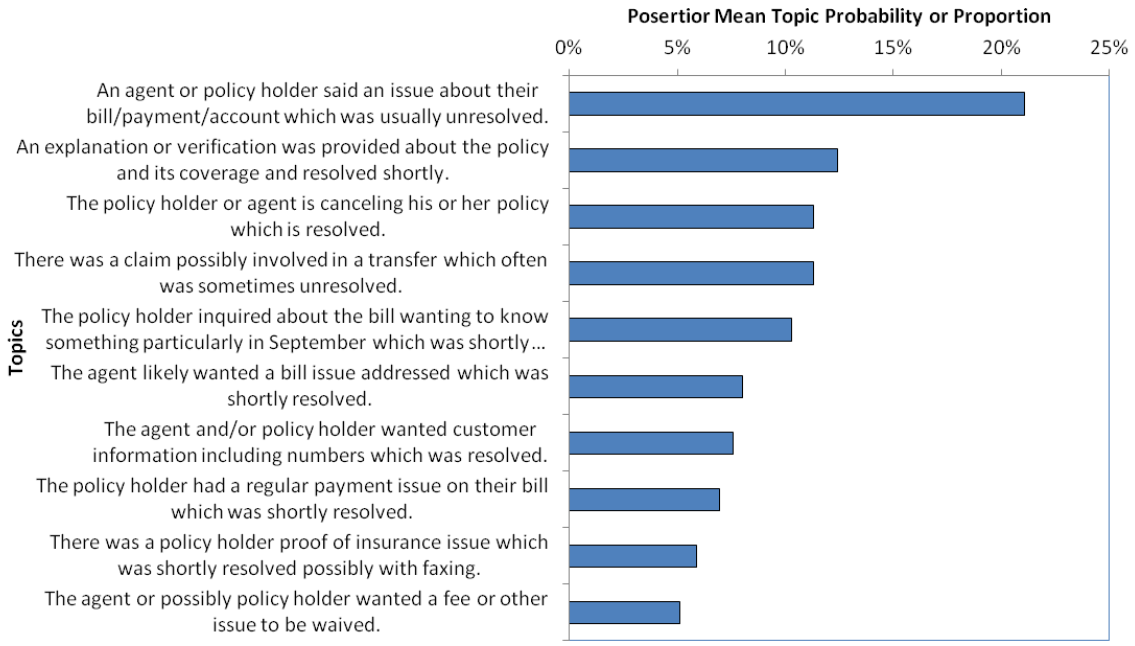

Figure 12. Call center clusters from SMERT model with manually entered interpretations.

Mulaik (1985) argued that iterative interpretation of salient features is often crucial in exploring problems. The analysis in Example 3, an approximate model of sentiment was first generated using recall counts. Then, inspecting the residuals, an additional factor relating to miles per gallon was hypothesized. Further investigation suggested that the effects for the F-150 associated with gas mileage might only be operating in the early years of the time period studied.

4.7 Final Remarks

In this article, we describe how the exploratory (EDA) framework of De Mast and Trip (2007) applies to text data. The resulting exploratory text data analysis ETDA principles are developed using examples from real-world quality improvement projects. In some sense, ETDA is an extension of EDA since there is an initial "preprocessing" step which could involve clustering, sentiment analysis, or another procedure which transforms the text into quantitative inputs for further analysis. We use the remainder of this section to review and expand on comments raised in De Mast and Trip (2007).

While automated algorithms could help in certain steps such as text data analysis and display (*Step 2*) and identification of salient features (*Step 3*), it is difficult to imagine that interpretation (*Step 4*) could easily be automated. In our examples, it required intuition to relate the topics or semantic relationships with possible causes of interest to practitioners. Automatic preprocessing, however, is perhaps the main motivation for the use of text modeling methods with millions of Tweets, e.g., being transformed in seconds into a Pareto chart as in Figure 8.

Also, it should be noted that ETDA only generates hypotheses and not confirmed results. Additional data collection and CDA are generally needed to generate facts about the causes of problems. The subjectivity of text, clustering, and semantic analyses only compound the inherent indeterminacy of EDA. Therefore, if the results of ordinary EDA are regarded skeptically, this skepticism should likely be deepened for EDTA.

Data available for ETDA is likely limited. First, the data might not be representative of the relevant populations. In Example 2, inputs from Consumer Reports

members may not be representative of the owner population. Second, while there is no clear problem with using the same data to generate and test hypotheses, the subjectivity of text data suggest an additional burden in collecting new data for confirmation will often be needed. Text data might rarely seem appropriate for proving physical effects in a manner like other types of engineering data.

We developed ETDA with various forms of text inputs to quality and design engineering in mind: surveys, complaint transcripts, customer ratings or maintenance squawks. We hope that the principles, methods, and diagrams introduced here may become a standard part of the analysis process for these types of data. Then, more promising hypotheses about the causes for quality problems and avenues for improvements may be generated in part because the clinical "mind set" commonly in use relating to other types of data can be extended to text data.

4.8 References

Allen, T.T. and Xiong, H. (2012). "Pareto Charting Using Multifield Freestyle Text Data Applied to Toyota Camry User Reviews". *Applied Stochastic Models in Business and Industry*, 28(2), pp. 152-163.

Allen, T. T.; Sui, Z.; and Akbari, K. (under 1st review). "Exploratory Text Data Analysis for Quality Hypothesis Generation". Submitted to *Journal of Quality Technology*.

Allen, T.T.; Xiong, H.; and Afful-Dadzie A. (2016). "A Directed Topic Model Applied to Call Center Improvement". *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Bisgaard, S. (1996). “The Importance of Graphics in Problem Solving and Detective Work”. *Quality Engineering* 9(1), pp. 157-162.

Blei, D.M.; Ng, A.Y.; and Jordan, M.I. (2003). “Latent Dirichlet Allocation”. *Journal of Machine Learning Research*, 3, pp. 993-1022.

De Mast, J. and Bergman, M. (2006). “Hypothesis Generation in Quality Improvement Projects: Approaches for Exploratory Studies”. *Quality and Reliability Engineering International* 22(7), pp. 839-850.

De Mast, J. and Trip, A. (2007). “Exploratory Data Analysis in Quality-Improvement Projects”. *Journal of Quality Technology*, 39(4), pp. 301-311.

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Good, I. J. (1983). “The Philosophy of Exploratory Data Analysis”. *Philosophy of Science*, 50, pp. 283-295.

Griffiths, T.L. and Steyvers, M. (2004). “Finding Scientific Topics”. *Proceedings of the National Academy of Sciences* 101(suppl. 1), pp. 5228-5235.

Hoaglin, D. C.; Mosteller, F.; and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, NY.

Liu, B. (2012). “Sentiment Analysis and Opinion Mining”. *Synthesis Lectures on Human language Technologies*, 5(1), pp. 1-167.

Mulaik, S. A. (1985). “Exploratory Statistics and Empiricism”. *Philosophy of Science*, 52, pp. 410-430.

Niiniluoto, I. (1999). “Defending Abduction”. *Philosophy of Science* 66 (supplemental), pp. S436–S451.

Pang, B. and Lee, L. (2008). “Opinion Mining and Sentiment Analysis”. *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1-135.

Porter, M.F. (1980). “An Algorithm for Suffix Stripping”. *Program* 14(3), pp. 130-137.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand Reinhold, Princeton.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. The Graduate School of the Department of Agriculture, Washington [reprinted by Dover Publications, New York, 1986].

Teh, Y.W.; Newman, D.; and Welling, M. (2006). “A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation”. In *Advances in Neural Information Processing Systems*, pp. 1353-1360.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, PA.

Turney, P.D. (2002) “Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews”. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Introduction

This chapter reviews the problem statements from the introduction chapter and summarizes the answers based on previous chapters. Directions for future research will also be discussed.

5.2 Answers to Problem Statements

1. How can we leverage social media data to support decision analyses unrelated to the social media?

To address this issue, Chapter 2 used “topic models” (e.g. Blei, Ng, and Jordan, 2003) to retrieve social media text data on specific subjects and, through minimal human inspections, to convert this into quantitative data available for Bayesian probability updates. Then, used these updates to support routine decision analyses, i.e., decisions of policy that are made repeatedly in different time periods.

2. How can these methods be computationally efficient for corpora involving tens of thousands of documents? How to make the results of these estimation methods repeatable and stable?

To address this issue, Chapter 2 explored the use of social media as an observation source for timely decision-making. To efficiently generate the observations for Bayesian updates, the dissertation proposes a novel computational method to fit an existing clustering model. It employed the concept of transforming k-means clustering results to estimate topic model parameters. The proposed method is called K-means Latent Dirichlet Allocation (KLDA). The computational results showed a promising result that KLDA is a computationally efficient method to fit approximate topic models with improved repeatability. The method is illustrated using a cyber security problem relating to changing maintenance policies during periods of elevated risk.

3. How could we leverage the models to identify an emerging topic?

Chapter 3 is mainly motivated by cyber security applications. In this chapter, a visual summarizing technique based on topic models and Twitter feeds is proposed to support passive summarizing and sensemaking. The associated "importance score" point system is intended to mitigate the topic models' weakness on identifying emerging topics. The proposed method is called TWitter Importance Score Topic (TWIST) summarizing method. TWIST employs the topic proportion outputs of tweets and assigns importance points to present trending topics. TWIST could generate a chart showing the important and trending topics that are discussed over a given time period. Two cyber security cases were tested against TWIST to test whether the method is sensitive to the emerging topic signals.

4. How will quality engineers and decision analysts deal with text data, visualization and analysis?

Chapter 4 proposed a general framework on how to work with text data to generate quality hypothesis. As a special case of Exploratory Data Analysis (EDA), Exploratory Text Data Analysis (ETDA) implements text as the input data and the goal is to extract useful information from the text inputs for exploration of potential problems and causal effects. Four major steps of ETDA in the quality improvement projects: pre-processing text data, text data processing and display, salient feature identification, and salient feature interpretation were explored alongside various case studies.

5.3 Future Work Opportunities

In this section, we suggest following opportunities for future work.

1. For the decisions problems in this dissertation, the current state selection may depend on previous states can potentially be investigated using Partially Observable Markov Decision Process (POMDP) formulations. Other techniques besides k-means-based estimation such as fuzzy c clustering, variational method and frequentist method can be explored. Also, additional comparison metrics and test cases might better clarify the accuracy limitations of KLDA methods. New evaluation metrics on accuracy could be more objective and interpretable than root mean square (RMS). Currently, the computational experiments involve only small test corpora from Allen, Xiong, and Afful-Dadzie (2016). Larger corpora from the literature can be explored. Additional visualization methods beyond Pareto charts can be tested as well as additional applications. Methods that

permit experts to edit topics offer the promise of more informative observations (Zhao et al. (2012); Sun (2014); Madsen et al. (2005); Allen et al. (2016); Sui et al. (2015)). The related methods can also be made more efficient using O(T), where T is the number of clusters or topics, estimation and related to decision problems.

2. TWIST can be compared with alternatives including methods based on more repeatable estimation procedures than collapsed Gibbs sampling. TWIST should also be made more automatic. Instead of including manually generated labels in Step 9 of TWIST, auto generation can be investigated. Also, TWIST based on the simpler LDA may be sufficient without human high-level data generation and the complications of SMERT. Moreover, the validation of TWIST could be explored with simulated numerical examples and the related statistical properties can be evaluated. Finally, domains outside of cyber security can be studied. These might relate, e.g., to sentiment analysis and the interests of populations relating to marketing or military conflicts.

3. To explore more applications of text analytics on decision problems, topic modeling on tweets related to finance should be explored on the application of trading financial derivatives. This will involve an application and extension of the Bayesian Adaptive Markov Decision Process (BAMDP) method. The trading instrument to investigate could be Bermudan option for S& P 100 index. It is the most liquid Bermudan option market available

worldwide (Kourtis and Markellos (2011)). The mean total market capital is about $120 billion (S&P 100 fact sheet (2016)). The Bermudan option gives the option holder the right to early exercise the option at pre-determined time points. Therefore, the decision on when to exercise the option is crucial in trading. Future research will use BAMDP to optimize exercise decisions policies by exploring multiple models or scenarios. The reason to use BAMDP is because to optimize considering parametric uncertainty in Markov decision processes, there are two major types of approaches: first, El Ghaoui and Nilim (2005) review robust methods seeking policies that are desirable for all parameters in an uncertainty set. But this method is conservative providing policies with limited status in decision analysis. For the second method, Delage and Mannor (2010) formulate the process as "data-driven Markov decision processes" (DDMDP), which adds an expectation over parametric uncertainty to the usual expectation over the intrinsic uncertainty. But Delage and Mannor (2010) only considered fixed policies which do not benefit from updating under new information. Therefore, Bayesian Adaptive Markov Decision Process (BAMDP) proposed by Duff (2002) is employed to model as a partially observable Markov decision process which promises seamlessly exploration and exploitation while always being expected profit optimal. Previously, BAMDP was applied to the cyber security field on helping making investment decision of cyber maintenance. But can BAMDP be applied to

the field other than the cyber security field? So future research will explore the possibility of the application of financial instrument trading by fitting the BAMDP model to see whether the model could achieve the optimal decision policy for exercising exotic options. Actions to be done to formulate the equity derivatives trading problem as a BAMDP with the following steps:

a. Generate actions for possible states and scenarios.
b. Generate of uncertain scenarios based on historical data and news events.
c. Generate of transition probability matrices and observation matrices from data.
d. Explore ways to implement BAMDP for finite horizon problems
e. Compare with other methods such as Longstaff and Schwartz method and comment on the benefits of the text data-driven implementation.

4. For the topic modeling, future work also includes adding the feature of synonym detection. Currently, the SMERT or k-means SMERT software only trim verbs or nouns of different forms into the word roots. But the software could not handle the situations of synonym. Therefore, future work should build a trained library to accommodate the new feature.
5. For ETDA, more methods of visualizing data should be explored. For example, Example 4 in Chapter 4, ratios or partial autocorrelation difference plots could be added. In order to remove the background noise, a baseline

topic could be assigned. Then the figure could present the ratio of one topic over the baseline topic. In this way, the background noises could be removed to present better signals or salient features.

6. To accommodate the applications in military, the sentiment analysis could be further developed with army-specific taxonomy. This is because the same word may represent different sentiment scores in different application fields. For example, “Code Blue” is very urgent and severe in the field healthcare, but may not be that important in other fields. Therefore, to further satisfy our needs in applications of military fields, the library or army-specific taxonomy should be built and trained.

5.4 References

Allen, T.T. and Hou, C. J. (preprint). “Optimal Learning Methods for Data-Driven Markov Decision Processes”. Working Paper, ISE, OSU.

Allen, T.T.; Xiong, H.; and Afful-Dadzie A. (2016). “A Directed Topic Model Applied to Call Center Improvement”. *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. (2003). “Latent Dirichlet Allocation”. *Journal of Machine Learning Research*, 3, pp. 993-1022.

Bok, H. S.; Kankanhalli, A.; Raman, K. S.; and Sambamurthy, V. (2012). “Revisiting Media Choice: a Behavioral Decision-Making Perspective”. *International Journal of e-Collaboration (IJeC)* 8(3), pp. 19-35.

Charalabidis, Y. and Loukis, E. (2012). “Participative Public Policy Making through Multiple Social Media Platforms Utilization”. *International Journal of Electronic Government Research (IJEGR)* 8(3), pp. 78-97.

Delage, E. and Mannor, S. (2010). “Percentile Optimization for Markov Decision Processes with Parameter Uncertainty”. *Operations Research* 58(1), pp. 203–213.

Duff, M. O. G. (2002). *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. Doctoral dissertation, University of Massachusetts Amherst.

El Ghaoui, L. and Nilim, A. (2005). “Robust Solutions to Markov Decision Problems with Uncertain Transition Matrices”. *Operations Research*, 53(5).

Gao, X.; Zhong, W.; and Mei, S. (2013). “Information Security Investment When Hackers Disseminate Knowledge”. *Decision Analysis* 10(4), pp. 352-368.

Gartner (2014). “Gartner Says Worldwide Information Security Spending Will Grow Almost 8 Percent in 2014 as Organizations Become More Threat-Aware”. Retrieved from http://www.gartner.com/newsroom/id/2828722 Accessed November 22, 2016.

Kourtis, A. and Markellos, R. N. (2011). “Traded American Options are Bermudan”. *Managerial Finance* 37(11), pp. 978 – 984

Miller, S.; Wagner, C.; Aickelin, U.; and Garibaldi, J. M. (2016). “Modelling Cyber-Security Experts' Decision Making Processes Using Aggregation Operators”. *Computers & Security* 62, pp. 229-245.

Parnell, G. S.; Butler III, R. E.; Wichmann, S. J.; Tedeschi, M.; and Merritt, D. (2015). “Air Force Cyberspace Investment Analysis”. *Decision Analysis* 12(2), pp. 81-95.

Paté-Cornell, M. E. (2012). “Games, Risks, and Analytics: Several Illustrative Cases Involving National Security and Management Situations”. *Decision Analysis* 9(2), pp. 186-203.

Smallwood, R. D. and Sondik, E. J. (1973). “The Optimal Control of Partially Observable Markov Processes over a Finite Horizon”. *Operations Research* 21(5), pp. 1071-1088.

Sui, Z.; Milam, D.; and Allen, T. T. (2015). “A Visual Summarizing Technique Based on Importance Score and Twitter Feeds”. *INFORMS Social Media Analytics Student Paper Competition*.

REFERENCES

Allen, T. T. and Hou, C. J. (preprint). "Optimal Learning Methods for Data-Driven Markov Decision Processes". Working Paper, ISE, OSU.

Allen, T. T. and Xiong, H. (2012). "Pareto Charting Using Multifield Freestyle Text Data Applied to Toyota Camry User Reviews". *Applied Stochastic Models in Business and Industry*, 28(2), pp. 152-163.

Allen, T. T.; Sui, Z.; and Akbari, K. (under 1st review). "Exploratory Text Data Analysis for Quality Hypothesis Generation". Submitted to *Journal of Quality Technology*.

Allen, T. T.; Sui, Z.; and Parker, N. (under 2nd review). "Timely Decision Analysis Enabled by Efficient Social Media Modeling". Submitted to *Decision Analysis*.

Allen, T. T.; Vinson, S. M.; Raqab, A.; and Alam, Y. (2013). "Using SMERT to Identify Actionable Topics in Student Feedback." *Integrated Systems Engineering Technical Report 2013*.

Allen, T. T.; Xiong, H.; and Afful-Dadzie A. (2016). "A Directed Topic Model Applied to Call Center Improvement". *Applied Stochastic Models in Business and Industry* 32(1), pp. 57-73.

Bisgaard, S. (1996). "The Importance of Graphics in Problem Solving and Detective Work". *Quality Engineering* 9(1), pp. 157-162.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3, pp. 993-1022.

Bok, H. S.; Kankanhalli, A.; Raman, K. S.; and Sambamurthy, V. (2012). "Revisiting Media Choice: a Behavioral Decision-Making Perspective". *International Journal of e-Collaboration (IJeC)* 8(3), pp. 19-35.

Borrero, J.S.; Prokopyev, O.A.; and Saur é, D. (2015). "Sequential Shortest Path Interdiction with Incomplete Information". *Decision Analysis* 13(1), pp. 68-98.

Bossenger, A. (2014). "How to Use Twitter Analytics to Find Important Data." *Social Media Examiner RSS*. N.p., 27 July 2014. Web. 05 May.

Britt, R. (2015). "How you and 'The Rock' Turned His Movie Around". Retrieved from http://www.marketwatch.com/story/how-hollywood-is-using-social-media-to-tell-if-a-movie-will-be-a-hit-Accessed June 19, 2015.

Cao, Y. (2014). "Reducing Interval-Valued Decision Trees to Conventional Ones: Comments on Decision Trees with Single And Multiple Interval-Valued Objectives". *Decision Analysis* 11(3), pp. 204-212.

Carpenter, B. (2010). "Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling". revision 1.4, LingPipe, Inc., carp@lingpipe.com, lingpipe.files.wordpress.com/2010/07/lda1.pdf (as of 5-4-2017).

Charalabidis, Y. and Loukis, E. (2012). "Participative Public Policy Making through Multiple Social Media Platforms Utilization". *International Journal of Electronic Government Research (IJEGR)* 8(3), pp. 78-97.

De Mast, J. and Bergman, M. (2006). “Hypothesis Generation in Quality Improvement Projects: Approaches for Exploratory Studies”. *Quality and Reliability Engineering International* 22(7), pp. 839-850.

De Mast, J. and Trip, A. (2007). “Exploratory Data Analysis in Quality-Improvement Projects”. *Journal of Quality Technology*, 39(4), pp. 301-311.

DeGroot, M. H. (2005). *Optimal Statistical Decisions.* Vol. 82, John Wiley & Sons (reprint 1970).

Delage, E. and Mannor, S. (2010). “Percentile Optimization for Markov Decision Processes with Parameter Uncertainty”. *Operations Research* 58(1), pp. 203–213.

Duff, M. O. G. (2002). *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. Doctoral dissertation, University of Massachusetts Amherst.

El Ghaoui, L. and Nilim, A. (2005). “Robust Solutions to Markov Decision Problems with Uncertain Transition Matrices”. *Operations Research*, 53(5).

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Followthehashtag. (2015). “Followthehashtag // Twitter Keyword Search Analytics, Influence, Geo Content Analysis Tool, and Much More.” N.p., n.d. Web. 04 May.

Gao, X.; Zhong, W.; and Mei, S. (2013). “Information Security Investment When Hackers Disseminate Knowledge”. *Decision Analysis* 10(4), pp. 352-368.

Gartner (2014). “Gartner Says Worldwide Information Security Spending Will Grow Almost 8 Percent in 2014 as Organizations Become More Threat-Aware”. Retrieved from http://www.gartner.com/newsroom/id/2828722 Accessed November 22, 2016.

Ghosh, S. and Dubey, S. K. (2013). “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms”. *International Journal of Advanced Computer Science and Applications* 4(4).

Good, I. J. (1983). “The Philosophy of Exploratory Data Analysis”. *Philosophy of Science,* 50, pp. 283-295.

Griffiths, T. L. and Steyvers, M. (2004). “Finding Scientific Topics”. *Proceedings of the National Academy of Sciences* 101(suppl 1), pp. 5228-5235.

Hoaglin, D. C.; Mosteller, F.; and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, NY.

Kourtis, A. and Markellos, R. N. (2011). “Traded American Options are Bermudan”. *Managerial Finance* 37(11), pp. 978 – 984

Lee, S. H. (2012). *Comparison and Application of Probabilistic Clustering Methods for System Improvement Prioritization*. Doctoral dissertation, The Ohio State University.

Liu, B. (2012). “Sentiment Analysis and Opinion Mining”. *Synthesis Lectures on Human language Technologies*, 5(1), pp. 1-167.

Lloyd, S. P. (1982). “Least Squares Quantization in PCM”. *IEEE Transactions on Information Theory* Vol. 28, pp. 129–137

Madsen, R. E.; Kauchak, D.; and Elkan, C. (2005). "Modeling Word Burstiness Using the Dirichlet Distribution". In *Proceedings of the 22nd International Conference on Machine Learning*. pp. 545-552.

Miller, S.; Wagner, C.; Aickelin, U.; and Garibaldi, J. M. (2016). "Modelling Cyber-Security Experts' Decision Making Processes Using Aggregation Operators". *Computers & Security* 62, pp. 229-245.

Moujahid, A. (2015). "An Introduction to Text Mining Using Twitter Streaming API and Python". *Data Analytics and More*. N.p., n.d. Web. 04 May.

Mulaik, S. A. (1985). "Exploratory Statistics and Empiricism". *Philosophy of Science*, 52, pp. 410-430.

Niiniluoto, I. (1999). "Defending Abduction". *Philosophy of Science* 66 (supplemental), pp. S436–S451.

Packiam, R. M. and Prakash, V. S. J. (2015). "An Empirical Study on Text Analytics in Big Data". In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. pp. 1-4.

Pang, B. and Lee, L. (2008). "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1-135.

Parnell, G. S.; Butler III, R. E.; Wichmann, S. J.; Tedeschi, M,; and Merritt, D. (2015). "Air Force Cyberspace Investment Analysis". *Decision Analysis* 12(2), pp. 81-95.

Paté-Cornell, M. E. (2012). "Games, Risks, and Analytics: Several Illustrative Cases Involving National Security and Management Situations". *Decision Analysis* 9(2), pp. 186-203.

Porter, M.F. (1980). “An Algorithm for Suffix Stripping”. *Program* 14(3), pp. 130-137.

Russel, M. (2015). “Mining the Social Web”. *Google Books*. N.p., n.d. Web. 04.

Russell, M. A. and Russell M. (2011). *21 Recipes for Mining Twitter*. O'Reilly Media, Inc.

Sent, D. (2015). “Python Programming Tutorials”. *Python Programming Tutorials*. N.p., n.d. Web. 05 May.

Shah, D. and Zaman, T. (2010). “Community Detection in Networks: the Leader-Follower Algorithm”. *arXiv preprint arXiv:1011.0774*.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand Reinhold, Princeton.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. The Graduate School of the Department of Agriculture, Washington [reprinted by Dover Publications, New York, 1986].

Smallwood, R. D. and Sondik, E. J. (1973). “The Optimal Control of Partially Observable Markov Processes over a Finite Horizon”. *Operations Research* 21(5), pp. 1071-1088.

Steyvers, M. and Griffiths, T. (2007). “Probabilistic Topic Models”. *Handbook of Latent Semantic Analysis* 427(7), pp. 424-440.

Sui, Z.; McCormick, C.; Allen, T. T.; and Milam, D. (2015). “Benchmarking Comparison of Methods Used to Extract Data From Twitter”. *INFORMS Social Media Analytics Student Paper Competition*.

Sui, Z.; Milam, D.; Allen, T. T. (2015). “A Visual Summarizing Technique Based on Importance Score and Twitter Feeds”. *INFORMS Social Media Analytics Student Paper Competition*.

Sun, X. (2014). “Textual Document Clustering Using Topic Models”. In *Semantics, Knowledge and Grids (SKG), 2014 10th International Conference on*. pp. 1-4.

Teh, Y.W.; Newman, D.; and Welling, M. (2006). “A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation”. In *Advances in Neural Information Processing Systems*, pp. 1353-1360.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, PA.

Turney, P.D. (2002) “Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews”. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424.

Von Neumann, J. and Morgenstern, O. (2007). *Theory Of Games And Economic Behavior*. Princeton University Press (2nd ed. 1947).

Yang, J. and Counts, S. (2010). “Predicting the Speed, Scale, and Range of Information Diffusion in Twitter”. *ICWSM* 10, pp. 355-358.

Zaman, T. R.; Herbrich, R.; Gael, J. V.; and Stern, D. (2010). “Predicting Information Spreading in Twitter”. *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS* 104 (45), pp. 17599-601.

Zaman, T.; Fox, E. B.; and Bradlow, E. T. (2014). “A Bayesian Approach for Predicting the Popularity of Tweets”. *The Annals of Applied Statistics* 8, pp. 1583-1611.

Zhao, T.; Li, C.; Li, M.; Wang, S.; Ding, Q.; Li, L. (2012). “Predicting Best Responder in Community Question Answering Using Topic Model Method”. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology* 01, pp. 457-461.

APPENDIX A. TOPIC MODELING CASE STUDY

This appendix contains data for the case studies including the true model which originally appeared in Allen, Xiong, and Afful-Dadzie (2015)

Table 6. Synthetic data for the numerical example

Doc#	Document
1	**The operator cut aluminum and dropped it at station1.**
2	**The inspector drilled plastic and overheated it at station2.**
3	**The manager milled steel and misaligned it at station3.**
4	**The engineer saw stone and over torqued on the truck.**
5	**The supplier welded and misdimensioned the titanium offsite.**
6	**The inspector drilled plastic and overheated it at station2.**
7	**It was drilled and overheated.**
8	**It was drilled and overheated.**
9	**The engineer and the manager at station3 and on the truck.**
10	**The welded titanium was misdimensioned.**
11	**The titanium was welded and misdimensioned offsite.**
12	**The steel was misdimensioned.**
13	**The operator cut the steel and plastic.**
14	**The manager welded it and misdimensioned it.**
15	**The operator cut and dropped the aluminum at station1.**
16	**The operator cut and dropped it at station1.**
17	**The engineer welded and misdimensioned the titanium.**
18	**It was drilled and overheated.**
19	**It was drilled and overheated.**
20	**The manager milled steel and misaligned it at station3.**
21	**The operator cut and dropped the steel at station1.**
22	**The engineer and the manager at station3 and offsite.**
23	**It was drilled and overheated.**
24	**The engineer saw stone and over torqued on the truck.**
25	**The stone was drilled and overheated.**
26	**It was drilled and overheated.**
27	**It was drilled and overheated.**
28	**It was drilled and overheated offsite.**
29	**The supplier welded titanium and misdimensioned it offsite.**
30	**The operator cut and dropped the titanium at station1.**
31	**The operator cut and dropped it at station1.**
32	**It was steel.**
33	**The steel was drilled and overheated.**
34	**It was drilled and overheated at station3.**
35	**The engineer and the manager at station1 and on the truck.**
36	**The welded titanium was misdimensioned.**
37	**It was drilled and overheated.**
38	**It was drilled and overheated.**
39	**The supplier welded titanium and misdimensioned it offsite.**
40	**It was drilled and overheated.**

Table 7. True model for the numerical example

T1	**0.4**	**T2**	**0.2**	**T3**	**0.15**	**T4**	**0.125**	**T5**	**0.125**
Word	Prob	Word	Prob	Word	Prob	Word	Prob	Word	Prob
Oper	0	oper	0	oper	0.23	oper	0	oper	0
Cut	0	cut	0	cut	0.23	cut	0	cut	0
alumi num	0	alumin um	0	alumin um	0.08	alumin um	0	alumin um	0
Drop	0	drop	0	drop	0.23	drop	0	drop	0
statio n1	0	station 1	0	station 1	0.23	station 1	0	station 1	0
inspec tor	0.1	inspec tor	0	inspec tor	0	inspec tor	0	inspec tor	0
Drill	0.35	drill	0	drill	0	drill	0	drill	0
plastic	0.1	plastic	0	plastic	0	plastic	0	plastic	0.1
overh	0.35	overh	0	overh	0	overh	0	overh	0
statio n2	0.1	station 2	0	station 2	0	station 2	0	station 2	0
mana g	0	manag	0	manag	0	manag	0.25	manag	0.1
mill	0	mill	0	mill	0	mill	0	mill	0.1
steel	0	steel	0	steel	0	steel	0	steel	0.5
misali gn	0	misali gn	0	misali gn	0	misali gn	0	misali gn	0.1
statio n3	0	station 3	0	station 3	0	station 3	0.25	station 3	0.1
engin	0	engin	0	engin	0	engin	0.25	engin	0
saw	0	saw	0	saw	0	saw	0	saw	0

Continued

Table 7 continued

stone	**0**	**stone**	**0**	**stone**	**0**	**stone**	**0**	**stone**	**0**
overto rqu	0	overto rqu	0	overto rqu	0	overto rqu	0	overto rqu	0
truck	0	truck	0	truck	0	truck	0.25	truck	0
suppli er	0	suppli er	0.05	suppli er	0	suppli er	0	suppli er	0
weld	0	weld	0.3	weld	0	weld	0	weld	0
misdi mens	0	misdi mens	0.3	misdi mens	0	misdi mens	0	misdi mens	0
titani um	0	titaniu m	0.3	titaniu m	0	titaniu m	0	titaniu m	0
offsit	0	offsit	0.05	offsit	0	offsit	0	offsit	0

APPENDIX B. TWITTER DATA EXTRACTION METHODS

B.1. Methods

B.1.1. Twitter Analytics

The Twitter analytics dashboard is an add on for users with advertiser status. These users can find detailed information on how their outgoing tweets are performing based on a few different criteria. The Twitter analytics tool provides public data only. If a user account has privacy settings and they do not follow the advertisers, their data will not be provided. This makes retrospective analyses for the previously unfollowed (apparently) impossible.

The dashboard is an intuitive tool for social media marketers. On the dashboard the user can see a maximum date range of 91 days of their past tweets performance. The dashboard shows the user the top ten accounts their followers follow, ranked by percentage. This information can be used to better understand what kind of information your followers are interested in on Twitter.

The data provided by Twitter includes a tweets impressions, link clicks, retweets, detail expands, favorites, embedded media clicks, user profile clicks, and replies. An impression is the number of times the tweet is read . Link clicks refers to the number of

times the URL was clicked. Detail expands is the number of times the tweet was clicked on the view details. A graph displays the past month of data, if a user wants to compare different months the data can be downloaded to a CSV file. This information would help the user determine which of their tweets was most effective in reaching their audience, or whether a certain time was most effective. There is also a feature that tracks follower increases or decreases and information on follower's location and gender. Tracking the audience's interests is a key feature of this method.

B.1.2. Follow the Hashtag

Analytics application called "Follow the Hashtag" has many useful features. The main dashboard section includes: total tweets, total impressions, total potential tweet impressions in followers timelines, and results from multiplying each contributor's keyword repetitions and it ś followers number and adding all contributors potential impressions. Other outputs include the total audience, total potential audience, result of adding each contributor followers number, impressions / audience, and impressions per user of a tweet with searched keyword.

Algorithms are utilized to analyze Twitter contributors' gender and percentage of males and females. Follow the hashtag allows users to export data to a PDF or CSV. Both will produce detailed sheets including: summary, top tweets, top users, gender, reach, blob chart data, geolocation, and stream data (tweet content, country etc.). The reports also include the best hour of the day, best day of the week for a hashtag's performance. An influence section is also included showing user by keyword repeats, top users by

influence score, and top users by keyword. An influence score shows the largest contributors in a searched keyword. There is a historical data feature where the user can recover tweets only up to 60 days old. A Twitter picture analysis is available which shows all the pictures related to a Twitter search, useful for picture based Twitter contests, or to get a general overview. An aggregated key repeats chart shows aggregated repetition values over time of the most repeated words related to your search. This chart shows the evolution of each related keyword discovering how a searched keyword is related to others over time. Overall, the usability and built in analytics shine but the extraction is somewhat limited.

B.1.3. Python plus Tweepy

Python is a widely used programming language that is easy to use, and effective for text analytics. In particular, it is used in multiple software programs for exporting data from Twitter. Tweepy is a python library that uses Twitter's application programming interface (API) to access public data. An API is a way for other programs to enter a given program. To access Twitter's streaming API, the user must create an app on Twitter. Once the Twitter API access is granted the user needs the API key. This API must be secret with an Access token and Access token secret. A script file is utilized to access live tweets, the file can search for specific keywords or usernames. The file can be run for any length of time depending upon the amount of data the user wishes to collect. This program can only pull current, live tweets, no historical data can be provided. The data can be exported to a CSV using a line of code. In our experiments we found that the

interface was difficult to understand and the derived CSV was somewhat miss-parsed, i.e., the fields were not cleanly usable in all cases.

B.1.4. Next Analytics

Next Analytics is a paid software program used for video and social media analytics, including Google, Facebook, Twitter, Instagram, and YouTube. Here, we focus on Twitter analytics. Next Analytics for Twitter is primarily associated with Microsoft Excel as an add-in function. Next Analytics could extract all the Tweets going back (apparently) to the start of Twitter itself. Users can select to extract tweets from their own account, followers' accounts, or any specific account. Yet, the extraction is on an account basis unlike, e.g., Twitter analytics. The output from Next Analytics is also in the format of excel including the information account name, tweet text, post time, account favorite number, account friend number, account follower number, and retweet number. The formatting of the output is well done so that not much effort is needed to clean up after the extraction before analysis.

B.2. Benchmarking

B.2.1. Criteria

The first criteria relates to which sources of tweets can be tracked. Message sources show where the data is from. For both Python plus Tweepy and Follow the Hashtag application methods, when the searching keywords are inputted, the two methods will search for the tweets with the keyword across the world. For the Twitter Analytics and Next Analytics methods, the data and analyses will only be from the activities of the user account. Moreover, Next Analytics also has the function for users to choose the data sources: account users themselves, followers, friends, or even specific accounts the users want to analyze. Therefore, if the analysis is targeted to specific users, the Twitter Analytics and Next Analytics methods should be applied.

Second, the analysis duration gives out the time range of data. For the Python plus Tweepy method, the data is real time. The program will search for and output the Tweets with the searching keywords from the time point of the start of the program until the stop command is inputted. As long as there is a tweet with the keyword posted online, the program will output it instantaneously. For analysis duration, Twitter Analytics and Follow the Hashtag application methods analyze the data in history for the time range specified by users. The time range can look back for up to two years. For Next Analytics, the software could extract data back to whenever Twitter keeps for the account holders.

Third, some extraction methods show the individual tweets and others do not. All four extraction methods permit the outputting of individual messages but Follow the Hashtag emphasizes the statistics and meta data making viewing individual tweets less direct.

All of the software permit significant customization and flexibility. The fourth criteria relates to the level of detail of derived outputs as subjectively assessed in our testing. We find that Follow the Hashtag and Python + Tweepy offer relatively sophisticated output, far beyond extracting the tweets themselves. The fifth relates to the ease for which summary statistics about the tweets can be obtained. These are the numerical portions of the outputs.

Again, some of the software such as Follow the Hashtag permit the easy derivation of detailed statistics. The sixth criteria is our subjective assessment of user friendliness. Here, we find that the programming environment is significantly less friendly than the others which have fairly standard graphical user interfaces. Finally, we also subjectively assessed how easy it is to derive outputs of various formats. Again, all of the software permits significant customization. Yet, we focus on emphasis and ease and find much greater potential for Python + Tweepy than the others.

B.2.2. Comparison

In this section, the four different extraction methods are compared using seven criteria. The results are shown in Table 8. Of all of the criteria, the ability to extract tweets historically (criteria 2) is the most important in our applications. Therefore, Next Analytics shines for our needs. Also, we ourselves are capable of producing summary statistics so the strength of Follow the Hashtag is less relevant. The extreme potential for customization of Python + Tweepy also makes that software relevant for consideration.

Table 8. Comparison Matrix

	Twitter Analytics	**Follow the Hashtag**	**Python + Tweepy**	**Next Analytics**
Message Sources	Users/followers	Whole network	Whole Network	Users/followers/friends /specific accounts
Analysis Duration	History (relatively limited apparently)	History (relatively limited apparently)	Real time	History (can go back for multiple years_
Displaying Message	Yes	Only partially	Yes	Yes
Output information detail level (message, like, forward)	Relatively limited	Extensive	Extensive	Relatively limited
Summary Statistics	Yes	Yes	No	No
User Friendly	Yes	Yes	No	Yes
Output format	Excel	Excel	Many txt, excel…	Excel

B.2.3. Applications and Industry Usage

In this section, we propose suggestions on how the various software might support activities in different industries. Each of the software programs has a strength, even Twitter Analytics might offer minimal installation. As Twitter is more and more popular as the means of communication and information publication, these alternative extraction methods could be used by different users to extract data and information for business needs.

Python plus Tweepy is a method extracting real-time original data and has a great ability handling large-scale data. Hence, this method could be applied to the cyber security industry. Hypothetically, high end users such as the Department of Defense (DoD) could apply this method to extract keywords posted on Twitter in real time. As long as the terrorists publish a tweet with dangerous messages, the DoD could summarize the account activity and take appropriate actions. Moreover, this method satisfies the business need for high-end news-based trading in high-frequency trading (HFT) on Wall Street. HFT traders could use Python plus Tweepy to track company names, key words, and trading news on Twitter at any given time. For example, the Wall Street Journal posts a message that profit of Google this year goes up. The trader will use the Python plus Tweepy method to track the Wall Street Journal Twitter account and "Google's profit goes up" information. Then, the Tweepy software can extract information in real time and transfer it to another program which will identify the keywords and semantics and further process the information to output the command on buying Google stocks. All these processes are carried out within microseconds or even nanoseconds automatically on computers. Therefore, for the business objective of real-time information and fast processing on the raw data, this method is the best.

For the Twitter Analytics method, its message source is primarily from the activities of the user account and the data statistics that is available directly to the users. Therefore, the result could be directly used by marketing professionals. Like the advertisements on Gmail and Facebook accounts, marketing departments of retailing companies could use this method to get the posting message of specific customers and

doing further analysis with keywords. Then, the companies would know what products the customer may be interested in and target users to send the corresponding advertisements.

The Follow the Hashtag application method could also be applied in a similar way with more built-in statistical information but also more installation burden. As this application also provides the information about the best hour of the day, best day of the week for a hashtags performance, retailers could use this function to know when their product related keywords are most active. Then, they could target their advertising and sales forces on those active time points for better sales efficiency. Moreover, because its message source is from the whole Twitter network, media or fashion industries could benefit from it. The media or fashion industries could set the statistical analysis duration to be only within recent months. Then, the statistical analysis will give the most popular keywords in this period of time. The media or fashion industries could use the results to analyze news or fashion trends and in turn customize their business to accommodate the public's needs and trends.

For Next Analytics, because it has many functions analyzing Google, Facebook, Twitter, Instagram, and YouTube, it could be applied to a range in analytics for different media. Moreover, because it performs (apparently) the best when analyzing historical data with user-friendly interface, it is a good method in extracting and analyzing historical tweet text. Moreover, the output also includes the total retweet number and the number of retweets can be used as an important indicator in the prediction model for social events and changes. Therefore, if users want to use Twitter to analyze historical

data for cyber security industry, media or fashion industries, or even use the results to analyze whether a Hollywood movie will be a hit, Next Analytics method should be selected.

B.3. References

Bossenger, A. (2014). "How to Use Twitter Analytics to Find Important Data." *Social Media Examiner RSS*. N.p., 27 July 2014. Web. 05 May.

Followthehashtag. (2015). "Followthehashtag // Twitter Keyword Search Analytics, Influence, Geo Content Analysis Tool, and Much More." N.p., n.d. Web. 04 May.

Moujahid, A. (2015). "An Introduction to Text Mining Using Twitter Streaming API and Python". *Data Analytics and More*. N.p., n.d. Web. 04 May.

Russell, M. A. and Russell M. (2011). *21 Recipes for Mining Twitter.* O'Reilly Media, Inc.

Russel, M. (2015). "Mining the Social Web". *Google Books*. N.p., n.d. Web. 04.

Sent, D. (2015). "Python Programming Tutorials". *Python Programming Tutorials*. N.p., n.d. Web. 05 May.

Sui, Z.; McCormick, C.; Allen, T. T.; and Milam, D. (2015). "Benchmarking Comparison of Methods Used to Extract Data From Twitter". *INFORMS Social Media Analytics Student Paper Competition*.

Zaman, T. R.; Herbrich, R.; Gael, J. V.; and Stern, D. (2010). "Predicting Information Spreading in Twitter". *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS* 104 (45), pp. 17599-601.