NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

NATURAL LANGUAGE PROCESSING OF SHORT COMMENTS FROM UNITED STATES NAVY SURVEY DATA

by

Marvin P. Salonga

March 2020

Thesis Advisor:	Lyn R. Whitaker
Second Reader:	Christine Cairoli, OPNAV N1T

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE NATURAL LANGUAGE PROCESSING OF SHORT COMMENTS FROM UNITED STATES NAVY SURVEY DATA	**5. FUNDING NUMBERS**
6. AUTHOR(S) Marvin P. Salonga	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Invaluable information pertinent to decision making in naval planning and policy can be extracted from free response survey comments. However, processing survey comments for analysis can cost considerable time and funding depending on the methods used. We extend the work of Cairoli's 2017 Naval Postgraduate School thesis, "Categorization of Survey Text Utilizing Natural Language Processing," and demographic filtering to aid the Navy in analysis of short-answer, free response comments from Navy surveys. Furthermore, we adopt similar approaches of text analysis from Layug's 2018 Naval Postgraduate School thesis, "Extracting Major Topics from Survey Text Responses Using Natural Language Processing," in our efforts to discover meaningful topics. Through our newly modified text-mining methods, we aim to enhance the assignment process of survey comments to a particular topic. We apply our approach through analyzing comments from the Navy Exit Survey question, "Why are sailors leaving?" and the Navy Milestone Survey question, "What will make sailors stay on active duty?." Our exploration of text mining processes includes implementing lemmatization, discovering topics using the relevancy metric Latent Dirichlet Allocation (LDA), creating a new comment-to-topic assignment process, and developing a web-based application that illustrates these methods.

14. SUBJECT TERMS Navy surveys, free response comments, topic labels, comment labeling, natural language processing, text analysis, Latent Dirichlet Allocation, LDA	**15. NUMBER OF PAGES** 67
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

NATURAL LANGUAGE PROCESSING OF SHORT COMMENTS FROM UNITED STATES NAVY SURVEY DATA

Marvin P. Salonga
Lieutenant, United States Navy
BS, University of California, Berkeley, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2020

Approved by: Lyn R. Whitaker
Advisor

Christine Cairoli
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Invaluable information pertinent to decision making in naval planning and policy can be extracted from free response survey comments. However, processing survey comments for analysis can cost considerable time and funding depending on the methods used. We extend the work of Cairoli's 2017 Naval Postgraduate School thesis, "Categorization of Survey Text Utilizing Natural Language Processing," and demographic filtering to aid the Navy in analysis of short-answer, free response comments from Navy surveys. Furthermore, we adopt similar approaches of text analysis from Layug's 2018 Naval Postgraduate School thesis, "Extracting Major Topics from Survey Text Responses Using Natural Language Processing," in our efforts to discover meaningful topics. Through our newly modified text-mining methods, we aim to enhance the assignment process of survey comments to a particular topic. We apply our approach through analyzing comments from the Navy Exit Survey question, "Why are sailors leaving?" and the Navy Milestone Survey question, "What will make sailors stay on active duty?." Our exploration of text mining processes includes implementing lemmatization, discovering topics using the relevancy metric Latent Dirichlet Allocation (LDA), creating a new comment-to-topic assignment process, and developing a web-based application that illustrates these methods.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AC	Active Component
AFO	Absolute First Occurrence
BUPERS	Bureau of Naval Personnel
BUPERS-3	Military Community Management (BUPERS-3)
CIMS	Career Information Management System
CLS	Candidate Label Score
CTS	Candidate token score
CVSS	Career Viewpoint Surveys and Studies
C-WAY	Career Waypoints
DTM	Document Term Matrix
ESR	Electronic Service Record
EDLN	Estimated Date of Loss to the Navy
Freq	Frequency
FH	First Half
FTS	Full Time Support
LDA	Latent Dirichlet Allocation
LS	Label Size
MSR	Minimum Service Requirement
N13	Military Personnel Plans and Policy Division
NPC	Naval Personnel Command
NSIPS	Navy Standard Integrated Personnel System
OPNAV	Office of the Chief of Naval Operations
PE	Physiological Episode (Survey)
RC	Reference Commonness
RFO	Relative First Occurrence
POS	Parts of Speech
PRD	Projected Rotation Date
PCS	Permanent Change of Station

PTT	Partial Technical Terms
SEAOS	Soft End of Active Obligated Service
TAP	Transition Assistance Program
TT	Technical Term

EXECUTIVE SUMMARY

Manually processing survey comments for analysis expends considerable time and money. Invaluable information can be extracted from these comments that can be highly pertinent to decision making at all levels of military personnel planning and policy. Without the appropriate analytic tools, reading individual response comments to Navy surveys is inefficient in quickly aggregating meaningful information. Valuable insight from these comments can be essential in providing solutions to the many challenges our Navy is tasked to solve.

The original text processing methodology created by Cairoli (2017) utilizes a logistic regression model to create an initial descriptive label to each individual comment. These labels are then utilized to narrow down the categorization of comments into relevant topics. Layug (2018) branches out and explores methods such as preprocessing text independent of an external reference corpus, creating a more generalized model, and automating the topic discovery process. Cairoli (2017) and Layug (2018) base their work on that of Chuang, Manning, and Heer (2012a, 2012b). Following in their footsteps, our first goal is to find a set of possible 1- to 3-word labels for every comment. To accomplish this, comments are preprocessed and assigned a set of 1- to 3-word candidate tokens or labels derived from the comment. A Candidate Token Score (CTS), also known as a Candidate Label Score (CLS), is given to every individual candidate token (Cairoli, 2017). The CLS is a linear function of statistical and linguistic variables that help determine a label's potential for describing a comment (Cairoli, 2017). Candidate labels with the highest CLS are used to determine the appropriate candidate label for each comment.

The second step is for the analyst to construct a set of meaningful topics and corresponding "keywords" that describe those topics. A "Topic Bin Key" is used to assign comments to a few meaningful topics. This part is labor intensive and requires substantial analyst input (Cairoli, 2017). To modify and expand on the processes used by Cairoli (2017) and Layug (2018), we explore the implementation of lemmatization in place of stemming; utilizing relevancy and LDAvis (Sievert & Shirley, 2015) as another method

for formulating topic bins; and creating an R Shiny (Chang et al., 2018) application that can assist the survey analyst in selecting appropriate topics.

To validate our research, we apply our approach to the Navy Exit Survey question, "Why are sailors leaving?" and the Navy Milestone Survey question, "What will make sailors stay on active duty?" and compare our results to that of Cairoli (2017). From Cairoli's (2017) methodology on analyzing "reason leaving" comments, "30.4% of binned comments are exact matches to both experts, 38.1% match the top bin of at least one expert, and 64.9% match one of the top three bins by either of the experts" (Cairoli, 2017). However, Cairoli (2017) notes that both experts agreed on their first-choice bin for about 46.2% of the responses. Our methodology finds approximately 45.1% of the expert binned comments are exact matches to ours, 56.6% of our bins match the top bin of at least one expert, and 74.6% of our bins match one of the top three bins by either of the experts.

As for the stay results, Cairoli's (2017) methodology shows that "27.3% of our binned comments are exact matches to both experts, 45.5% match the top bin of at least one expert, and 56.8% match one of the top three responses by any of the experts with 43.2% of the primary ranked topic matching for the experts" (Cairoli, 2017). When implementing our methodology, our results shows that 25.0% of our binned comments are exact matches to both experts, 40.9% match the top bin of at least one expert, and 68.2% match one of the top three responses by any of the experts.

References

Cairoli, C. M. (2017). *Categorization of survey text utilizing natural language processing and demographic filtering.* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle.net/10945/56109

Chuang, J., Manning, C., & Heer, J. (2012a). *Termite: Visualization techniques for assessing textual topic models*, ACM. doi:10.1145/2254556.2254572

Chuang, J., Manning, C., & Heer, J. (2012b). Without the clutter of unimportant words. *ACM Transactions on Computer-Human Interaction (TOCHI), 19*(3), 1–29. doi:10.1145/2362364.2362367

Layug, C. (2018). *Extracting major topics from survey text responses using natural language processing* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle/10945/60425

Sievert, C., & Shirley, K. (2015). LDAvis: Interactive visualization of topic models. Retrieved from https://CRAN.R-project.org/package=LDAvis

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Lyn Whitaker, for her utmost patience, insight, and pedagogy, which have been highly appreciated throughout my academic endeavors at NPS. With your guidance I built my confidence in programming and also in developing my skills in a variety of areas within data analysis. When my family was going through a lot during my last year, you provided support and reminded me of how important time with family is and how limited it can be. I appreciate your wonderful character and class.

Additionally, I would like to thank my second reader, Christine Cairoli, for being so amazing and providing sound advice throughout the entire process. I am truly grateful for all the pep talks and words of wisdom that helped me stay motivated.

I would like to thank Daniel Diaz, Jan Lim, Alejandro Gonzales, Wil Vega, Nickos Leondaridis-Mena, Ji Jiang, Cang Pham, and Mansfield Murph for being members of my NPS family. I will never forget all of the times that we had; it was with you all that made my experiences in Operations Research so memorable and defining in my life.

Last, but not least. I would like to thank my mother, Elizabeth; my father, Morado; my aunt, Thelma; and my incredible sister, Marjorie, for being such a solid foundation in my life since day one. Through thick and thin. *Ai Ai Aten*.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Naval personnel surveys collect information concerning multiple aspects of naval service and employment. However, processing survey comments manually expends considerable time and money. Invaluable information can be extracted from these comments and can be highly pertinent to decision making at all levels of military personnel planning and policy. Cairoli (2017) and Layug (2018) developed text-mining processes based on surveys involving naval retention and aviation physiology, respectively; however, there is still room for improvement. We explore the application of new methods to both versions of the Navy Retention Survey (Navy Standard Integrated Personnel System, CVSS, n.d.) for a more recent objective analysis of free response comments to survey questions. An example question, from the Navy Milestone Survey, would be "What can be done to encourage you to remain in the Navy on active duty when you are next required to make a stay/leave decision?" This analysis facilitates guidance in finding factors that influence sailors' career intentions. Our modifications to Cairoli's (2017) and Layug's (2018) work include implementing lemmatization, analysis using a relevancy metric, creating a new binning process, and developing a web-based application to illustrate the methods for analyzing our data.

A. BACKGROUND

From a text-mining perspective, each text response to a survey question is treated as a "document." There are a variety of methods for identifying the major topics in a corpus of documents. Latent Dirichlet Allocation (LDA) is a "generative probabilistic model of a corpus with the basic idea that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" (Blei, Ng, & Jordan, 2003). According to Layug (2018), the Office of People Analytics successfully applied LDA analysis to a Department of Defense survey consisting of over 24,000 comments. However, it is important to note that most of the successful applications of topic models are for longer documents; whereas, smaller data sets of less than 1,000 short responses are shown to be less effective with certain types of LDA (Chuang et al., 2012b). The expected

number of responses that Naval personnel surveys have ranges from several thousands to only a couple hundred; furthermore, Navy survey comments tend to be short and are either too limited in quantity, are very domain specific or laced with technical terminology, slang, and misspellings that are difficult to understand if not read in context. Cairoli (2017) adapts a label selection approach of Chuang et al. (2012b) to address these limitations and applies it to survey comments of a smaller scale.

The comment analysis approach developed by Cairoli (2017) has two parts. The first part assigns a label to each text response or comment, almost completely automating the process. Every comment is cleaned according to set rules and tokenized into 1 to 3 consecutive-word combinations that are referred to as candidate tokens or candidate labels. Cairoli (2017) assigns two different types of variables to each candidate label, comment-specific variables and label-specific variables. These variables, such as the number of words in the label or label's size, help determine how well the candidate label describes a given comment. To assist in constructing these variables, Cairoli (2017) uses a reference corpus; the reference corpus augments the vocabulary of a survey comment corpus, establishing vocabulary that remains fairly consistent. Any relevant document specific to the survey topic can act as the reference corpus, which means selection can vary (Layug, 2018). A score which factors in both label specific and comment-specific variables is then computed for each candidate label; this score, the Candidate Token Score (CTS), is derived from a linear function of these variables utilizing estimated regression coefficients (Cairoli, 2017). The candidate label with the maximum CTS serves as the label for that comment. Cairoli (2017) and Layug (2018) use the term "Candidate Token Score" or "CTS." We will refer to this variable as the Candidate Label Score (CLS).

The first part of the process may yield hundreds of different labels, too many to use effectively. The second part of the process groups the comment labels into primary topic bins using a topic bin key that consists of keywords created using LDA and other graphical text-mining methods (Cairoli, 2017). These topic bins correspond to meaningful categories. In this step, the analyst uses subject matter experience for topic discovery and to sort the labeled comments into the appropriate topic bins. Classifying comments into a few meaningful topic bins allows the analyst to quantify results based on free text responses

to survey questions. Cairoli (2017) applies this methodology to analyze the questions "Why are sailors leaving?" and "What will make sailors stay on active duty?" posed in the Navy Exit Survey and Navy Milestone Survey, respectively. Cairoli's (2017) work on topic modeling enables future Navy Retention Survey analysts to sort comments by variables such as military rank, gender, and community. Moreover, Cairoli (2017) provides results that can influence leadership to modify or design policy incentives that retain our most capable sailors while also considering budget constraints and operational requirements.

Layug (2018) builds on Cairoli's (2017) work by implementing an approach that partitions a comment into several candidate items. The reasoning behind this is that comments may contain or address multiple topics. For example, in the Physiological Episode (PE) Survey examined by Layug (2018), the question "Based on your personal or second-hand knowledge of PEs, why do you think there has been a recent increase in reported episodes?" leads to responses with a numbered list of reasons, each of which may address a different topic. In our examples, we do not partition a comment into several comment items. Our objective is to improve the process of assigning a topic to a comment and thus we follow Cairoli (2017) in this aspect. It is important to note that, in our application, each comment label is assigned a unique identification code corresponding to the assigned comment. Thus, our methods can also apply to multiple–item comments.

One approach that we adopt from Layug (2018) is the omission of an external reference corpus. Cairoli (2017) uses a reference corpus to construct a measure of commonness, also known as Reference Commonness (RC). RC is a label-specific variable used in the CLS calculation, which can be a good indicator of a descriptive label (Cairoli, 2017). Another reason for using a reference corpus is to have one source determining if a candidate label is a partial technical term (PTT). We discuss technical terms (TT) and PTT thoroughly in Chapter II. Having a single source helps give a consistent definition of PTT over different questions in a given survey or over the same question administered in different surveys (Cairoli, 2017). However, Layug (2018) conducts an experiment comparing the use and omission of a reference corpus and concludes, although a reference corpus can be helpful for data sets less than 1000 responses, it may not be needed.

B. THESIS OUTCOMES

As Layug (2018) points out, unexpected results can arise when applying new approaches to different datasets. These findings can relay information on the various aspects of our methods that may need to be altered or revamped for efficiency. Through analyzing text comments from recent Navy Retention Survey data, we modify Cairoli's (2017) and Layug's (2018) approach in several ways. We implement the use of lemmatization, cross reference our results using a separate LDA model involving a relevancy metric, create a new binning process, as well as developing a web-based application using the R Shiny package (Chang et al., 2018) from the R statistical computing software (R Core Team, 2018) to aid the analyst in the topic binning portion of our processes.

1. Stemming versus Lemmatization

In essence, stemming reduces words into their respective word stem equivalent. For example, "chasing," "chases," and "chased" may all become "chas." An alternate way of looking at stemming is that it truncates a word to a form that is smaller or close to the same length depending on heuristic conditions. Stemming utilizes a heuristic measurement to determine these particular word stems and thus differentiates these stems from structural roots that are found within a dictionary. Lemmatization, on the other hand, returns the base dictionary form of a given word, known as the lemma, by conducting morphological analysis of words (Manning, Raghavan, & Schütze, 2008). With lemmatization for example, "chasing" becomes "chase." The results of stemming potentially make it difficult for researchers when conducting analysis in that there can be more room for misinterpretation. Whereas, with lemmatization, the results appear to be more clear since there are no partial words due to heuristic calculations.

2. LDA Relevancy Metric

An important implementation in our research is the relevancy metric, defined in Chapter II, and the use of the R package LDAvis (Sievert & Shirley, 2015). This package and the relevancy metric when used in our application, gives the analyst an interactive tool for visualizing potential topics and topic bin keys. This approach, along with other methods

outlined in Chapter II, enables the analyst to have a starting point in choosing an initial set of meaningful and relevant topics. Through several iterations of the processes outlined in Chapter II, the analyst refines the set of selected topics used to bin the comment labels accordingly.

3. Creating Topic Bins and Comment Label Assignment Process

A new method that we implement is an automation of assigning comment labels to topics. Once the analyst identifies an initial set of topics, each comment is assigned to a topic using the candidate label with the highest CLS. A comment that cannot be assigned to a topic using its label is placed in the *Other* category. Cairoli (2017) and Layug (2018) then analyze the entire text of the *Other* comments and manually assign them to a topic. Our new approach instead leverages the CLS scores. If we cannot assign a topic to a comment based on the candidate label with the highest CLS, we then consider the candidate label with the second highest CLS for that comment, and then the candidate label with the third highest CLS is used. Even so, the number of comments left over in the *Other* category may be large. Our R Shiny (Chang et al., 2018) application provides the analyst with tools adapted from Cairoli (2017) and Layug (2018) to identify additional topics to add to the initial topic list.

4. Web-Based Application

We develop the framework for a working R Shiny (Chang et al., 2018) application that guides the user through our methodology step by step. Our web-based application is intended for Navy survey analysts with a background in operations analysis, however, the web-based application is still a work in progress and can be adjusted to serve users with limited understanding of text mining and natural language processing.

C. THESIS OUTLINE

Chapter II covers our thesis methodology. Chapter III focuses on the background information and details of the Navy Retention Survey data utilized for our analysis along with examples that illustrate the methods utilized to obtain our results. Chapter IV delves into a deeper discussion of the methods that we employ along with the validation section

that compares the results of our approach to that of our Cairoli's (2017) methodology. Conclusions and recommendations for potential follow on research is given in Chapter V.

II. METHODOLOGY

The procedures we take to analyze our survey data adhere closely to the foundational work of Cairoli (2017), implement important modifications made by Layug (2018), and delve into other branches of natural language processing. The adaptations stem from approaching the issues that have arisen through the work of both Cairoli (2017) and Layug (2018). The explanation of our initial methodology is designed after that of Cairoli (2017) and Layug (2018). Section A of this chapter describes the steps (illustrated in Figure 1) taken to identify and rank comment labels for each survey response. Section B of this chapter details how we establish topic bins through the use of ordering the candidate labels we extract. Our work helps enhance the tool set available for future Navy Retention Survey analysts by automating portions of this process.

A. COMMENT LABELS

1. Standard Preprocessing

The preprocessing steps involve extracting punctuation with the exception of periods, commas, semicolons, colons, and punctuation that are necessary for parts of speech tagging (Cairoli, 2017). We then convert common contractions such as “I’m,” “you’re,” “we’re,” “they’re,” etc. to their corresponding whole words while also converting all text to lowercase. For now, we do not remove stop words like “a,” “the,” or “and” since survey responses are generally short and stop words may be necessary descriptors. Note, unlike Cairoli (2017), our examples in this thesis do not replace certain words using a predefined word dictionary. For these examples, we choose to not create a predefined dictionary as to limit the changes made to the raw text. However, using a predefined dictionary tailored to a specific survey, for example replacing “F-18,” “F18,” “F 18,” “F/A-18” with a common term, at times can be quite helpful in the analysis. Similar to both Cairoli (2017) and Layug (2018) we choose to not convert words to their root form through stemming.

2. Candidate Labels

Upon completion of standard preprocessing, every comment goes through a tokenization algorithm to find corresponding tokens. Tokens can be thought of as consecutive-word phrases or *n*-grams, which are any *n* consecutive words in a comment (Cairoli, 2017). For uniformity within this thesis, we refer to tokens as labels, but these terms can be used interchangeably. It is important to note that Chuang et al. (2012b) demonstrates that there is little added benefit to using more than three words in a label. With this in mind, we find all 1- to 3-grams for each comment (unigrams, bigrams, and trigrams). This allows us to define the set of candidate labels for each comment.

3. Candidate Label Score

Once each comment is broken down into 1- to 3-gram labels, we assign a score, the CLS, to each candidate label that serves to measure how well that label might serve as the main descriptive label for its comment. Chuang et al. (2012b) develops a method to find descriptive labels for larger text, and both Cairoli (2017) and Layug (2018) adapt the process to shorter text comments with character limits. The CLS is a linear function of statistical and linguistic variables that aids in discerning a label's potential for describing the comment (Cairoli, 2017). The two types of variables assigned to each candidate label are label-specific variables and comment-specific variables. These variables, discussed later in this chapter, factor into calculating CLS. We calculate a CLS for every unique candidate label; Cairoli (2017) and Layug (2018) take the label with the highest CLS as the label for that comment. In our research, we also retain candidate labels with the second and third highest CLS to be used when assigning comments to topics.

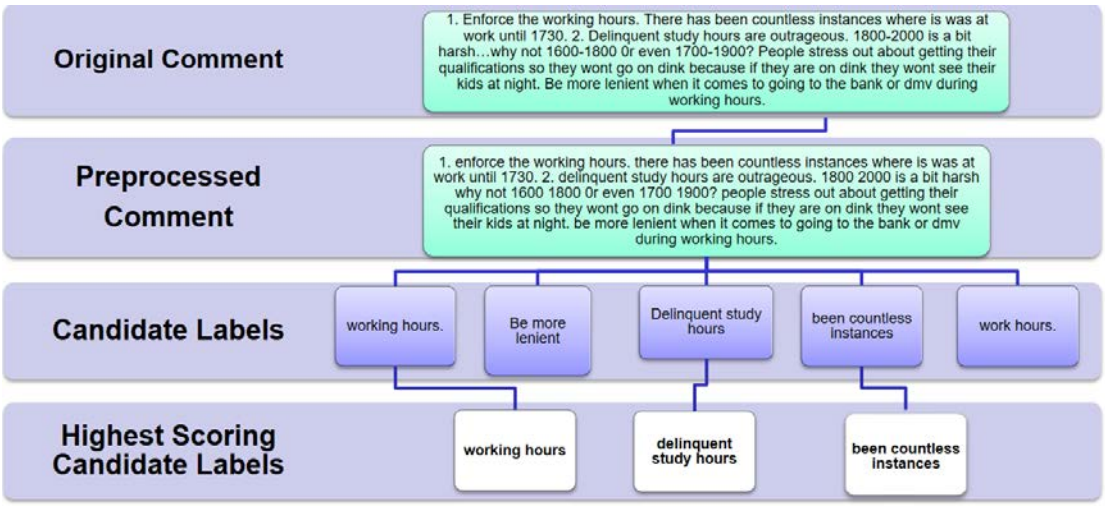

Figure 1. Candidate Label Scoring Process

a. Label-Specific Variables

Following Cairoli (2017) and Layug (2018), we define label-specific variables as being "unique to each candidate [label] and independent of the comment associated with the label. These variables are calculated once for each unique candidate label in the corpus and are then factored into the final [CLS] computation" (Cairoli, 2017). Our approach has three label-specific variables: Label Size (LS), Technical Term (TT), Partial Technical Term (PTT).

(1) Label Size

The LS variable counts the number of words in the label. In other words, it identifies if a candidate label is a unigram, bigram or trigram. Since many candidate labels begin with the same word and many contain substrings of other candidate labels, LS allows us to illustrate the impact of multi-word labels when it comes to describing comments (Cairoli, 2017).

(2) Technical Terms and Partial Technical Terms

TT represents a binary variable that specifies if a label is or is not a technical term. A TT can be defined in a variety of ways and thus has no generally accepted definition. Following Layug (2018) and Cairoli (2017), we will define TT as "a multi-word phrase

that meets a specific pattern; it begins with either an adjective or noun, strings together adjectives, nouns, or prepositions in the middle, and ends in a noun."

In the previous methodology, the variable "TT is assigned a value of 1 if a candidate label is found to match a TT or 0 otherwise" (Cairoli, 2017). The variable partial technical term (PTT) is defined as a substring of a given technical term. PTT is also a binary variable; it indicates whether a label is or is not a partial technical term. Like the TT variable, PTT will take the value 1 if the given candidate label is identified as a substring of a TT within the TT list and 0 otherwise. Following Layug's (2018) approach, we eliminate the use of an external reference corpus to construct the TT list. Our TT list is taken from the set of all candidate labels identified as TT. With the use of udpipe (Wijffels, 2018), a search for Parts of Speech (POS) patterns are conducted in each comment. Direct matches to POS patterns that are considered TTs are also identified. Further elaboration on the use of this approach will be detailed in Chapter IV.

b. Comment-Specific Variables

The other variables used for CLS computation are comment-specific variables. These variables differ from label-specific variables in that they capture how often and where the candidate label appears in each comment. Also, comment-specific variables vary from label-specific variables in that their values are comment dependent. Similar to Cairoli's (2017) work, we use the same three comment-specific variables: "Freq, the frequency of the [label] in each comment; [Relative First Occurrence] RFO, a measure of the first occurrence of a [label] relative to a [label] of the same frequency; [First Half] FH, an indication of whether a [label] is contained in the first half of a comment or not" (Cairoli, 2017).

(1) Frequency

We utilize both candidate labels and the comment corpus to build a document term matrix (DTM). The DTM is comprised of one row per comment (or in the case of multi-item comments in Layug's [2018] work, one row per item) and one column per candidate label; it stores the frequency count for that label by comment. In the DTM produced, the candidate label frequency count for each comment is greater than or equivalent to one. The

work of Chuang et al. (2012b) points out that candidate labels that have a higher frequency of appearance in the corpus are usually more important. Both Cairoli (2017) and Layug (2018) point out that candidate labels usually appear only once since survey comments are generally short in length; thus, the Freq variable may not have as great an impact on the importance of a candidate label for survey comments.

(2) Positional Elements

In the work of Chuang et al. (2012b), the candidate label position or location in reference to a document is practical in searching for descriptive labels. It is usually the case that when candidate labels appear frequently in the beginning of a document they tend to be more important than labels that appear frequently later in the document. We apply a similar approach to comments that Chuang et al. (2012b) uses for documents. We use absolute first occurrence (AFO) as a normalized measure ranging from 0 to 1 to represent the location of a label's first appearance within a given comment. Like Chuang et al. (2012b), we take the total number of words as well as taking the normalized position of the first word in the phrase to calculate AFO. Following Cairoli (2017), in the case of bigrams and trigrams, we consider the *n*-grams as a single "word." This allows us handle bigrams and trigrams that start in the same position to not share the exact same absolute first position. The one exception are the labels that begin the comment.

We use AFO to derive relative first occurrence (RFO). RFO is defined as a measure of "how likely a term is to initially appear earlier than a randomly-sampled phrase of the same frequency" (Chuang, 2013). Let k be the frequency of a label in the comment, then

$$RFO = (1 - AFO)^k. \tag{0.1}$$

Whether a label occurs in the first half of a comment is the next positional comparison we make. As aforementioned, Chuang et al. (2012b) point out that labels appearing first in a document (in our work "comment"), are generally more useful descriptors than labels positioned later in the document. We follow a similar approach, made by Cairoli (2017), to identify comment labels that appear in the first half of a

comment. We define FH as a binary variable that expresses whether a label is positioned within the comment's first half.

c. Candidate Label Score Calculation

Cairoli (2017) adopts the approach originated by Chuang et al. (2012b) to calculate a CLS by using estimated regression coefficients. Cairoli (2017) randomly samples two hundred comments composed of five or more words from four questions across three surveys to construct a dataset. A subject matter expert reads each of the 200 comments to determine the primary 1- to 3-gram label for every comment and stores them as the *expert label*. Cairoli (2017) then extracts all 1- to 3-gram candidate labels made from each survey comment excluding the *expert label*. Once this is done, ten of these are chosen randomly for each comment to act as incorrect labels assigned to that comment. This gives two types of labels: one correct, the expert label, and ten incorrect labels for each comment. Cairoli (2017) produces a 2,200 paired labels and comments data set, determines the variable values for each label comment pair, and estimates regression coefficients through fitting a logistic regression where the response variable is 1 if the label is the expert label and 0 otherwise.

With the exception of omitting the use of a reference corpus for determining PTT and RC, Layug (2018) calculates CLS using estimated regression coefficients in a similar manner. Layug (2018) picks 180 random comments from several surveys consisting of at least five words and by following the same steps as Cairoli (2018), produces a data set of 3,960 responses that include both expert and randomly chosen labels. Layug (2018) combines this data set with Cairoli's (2017) and calculates the comment and label-specific variable values for every label-comment pair. To calculate new regression coefficients, Layug (2018) fits a logistic regression in which the response variable is set to 1 whenever there is a comment label that matches an expert label; if a comment label is not matched to an expert label, it is set to 0. Our methodology utilizes the regression coefficients of Layug (2018), these coefficients are shown in Table 1. From Table 1, we see that LS is treated as a three-level categorical variable, rather than as a numeric variable.

Table 1. Regression Coefficients for Candidate Label Score Calculation. Adapted from Layug (2018).

Model Variable	Coefficient Estimate	Standard Error
(Intercept)	-1.310 ***	0.284
LS = 2	-0.882 ***	0.260
LS =3	-1.005 ***	0.278
TT	2.313 ***	0.321
PTT	-0.023	0.260
Log(Freq)	0.918 **	0.501
RFO	2.025 ***	0.405
FH	0.435 **	0.262

Statistical significance = ***: $p < 0.001$, **: $p < 0.1$

4. Primary Comment Label

Using the coefficients from Table 1, the candidate labels for each comment are scored accordingly. The primary comment label for a particular comment is the candidate label garnering the maximum CLS among the candidate labels for that comment.

B. GROUPING COMMENTS INTO TOPIC BINS

Finding meaningful topic bins and placing the comments into appropriate bins involves several methods. These methods rely heavily on the (primary) label chosen for each comment. Our first method begins with Layug's (2018) approach of constructing a list of potential terms or keywords. We create two reference lists of keywords that consist of the most frequent bigrams and the most frequent trigrams taken from the comment labels. The second approach utilizes an LDA model applied to a corpus of comment labels (where each comment's label serves as a document in the corpus) to create a third reference list of keywords that consists of the most salient 1- to 3-grams. Furthermore, the package LDAvis (Sievert & Shirley, 2015) is used in this second approach to help discern potential topics. The third method that we use is a correlation network that uses the most frequent and most salient 1- to 3-grams from the comment labels to construct a network evaluation

of correlation. These three methods help in formulating an initial table of topics and their respective keywords that aid in assigning comments to topics. Cairoli (2017) calls this list a "topic bin key." The details of all three methods are described in the following sections along with the algorithmic alterations we make.

1. Frequent Bigrams and Trigrams Reference Lists

We use a systematic approach similar to Layug (2018) that evaluates the comment label relationships to produce reference lists for finding meaningful topics and topic keywords. The first method adapted from Layug (2018) tokenizes the comment labels into bigrams and trigrams and omits those that contain stop words. However, there is an exception for permitting stop words and that is when a stop word occurs in the middle of a trigram (i.e., "availability of jobs," "options for transfer," and "lack of authority"). The bigrams and trigrams that remain are arranged by frequency. The thresholds for choosing the most frequent bigrams and trigrams are modified by the analyst until a reasonable list of keywords is produced.

2. LDA Saliency Reference List and Relevancy

The second keyword extraction method of Layug (2018) is more complex than the first, however it may find keywords that the first method did not. We implement the use of a corpus where each "document" is the comment label. We then tokenize the comment labels and design a DTM that consists of all possible 1- to 3-gram comment labels. Instead of utilizing stemming like Layug (2018) in this approach, we utilize lemmatization. Following Cairoli's (2017) methods and reasoning, we fit an LDA model to find inherent topics within our data. LDA topic modeling is useful to discern topics generated from our comment labels; LDA determines the latent topics T within a document corpus and enables a given document d to correspond several topics. From fitting an LDA model, we can estimate the word (or token) distribution for each topic, $\{p(w|T)\}$ where w indicates a word and the topic distribution for each document, $\{p(T|d)\}$ (Silge & Robinson, 2017). These estimated distributions facilitate the process of finding distinct, salient, and relevant words within the comment labels.

As defined by Chuang et al. (2012a), the distinctiveness of each token *w* uses the Kullback-Leibler divergence (Kullback & Leibler, 1951) and measures how much the conditional topic distribution, {*p*(*T*|*w*)} diverges from the unconditional topic distribution, {*p*(*T*)}. By using Bayes' rule to calculate {*p*(*T*|*w*)} and the label DTM for calculating {*p*(*T*)}, *distinctiveness* is defined as

$$distinctiveness(w) = \sum_{T} p(T \mid w) \log\left(\frac{p(T \mid w)}{p(T)}\right). \tag{0.2}$$

Chuang et al. (2012a) uses *saliency* to identify important, but not excessively frequent words that can be used to distinguish topics. Saliency can be described as the product of the probability of choosing a word *w* within a corpus of words, *p(w)* and the distinctiveness. The definition of a word's saliency is

$$saliency(w) = p(w) * distinctiveness(w). \tag{0.3}$$

To draw out each 1- to 3- gram with the highest saliency from the labels, we apply the fitted LDA model to our label DTM (Cairoli, 2017). We proceed to include new 1- to 3- grams extracted from this method into our keywords list in the topic bin key.

Although we do not use it to define keywords, another measurement that can be computed using LDA output is relevancy of a word w for a particular Topic T, as defined by Sievert and Shirley (2015). The definition of relevance given in Equation (2.4) is a weighted average with a weight parameter λ

$$relevancy(w \mid T) = \lambda \log\left(p(w \mid T)\right) + (1-\lambda)\log\left(p(w)\right). , \tag{0.4}$$

where $0 \le \lambda \le 1$. The use of relevancy will be detailed in our example that utilizes LDAvis (Sievert & Shirley, 2015) in Chapter IV.

3. Networks

The last method borrowed from Layug (2018) utilizes a network of unigrams originating from the comment labels. Layug (2018) creates a list containing the most frequent and most salient unigrams from an LDA fit of the label corpus and then calculates

the correlation between all unigram pairs. Layug (2018) points out that survey responses may begin by repeating the initial part of the question, which leads to many of those words being either labels or keywords. However, this may or may not provide the analyst with significant information. To account for these question words, Layug (2018) introduces a correlation threshold. The correlation for combinations that consist of question words must be greater than the threshold to be considered. We alter Layug's (2018) method by fitting an LDA model using all 1- to 3-grams taken from the comment label corpus. We also expand on the network implementation by utilizing different R packages such as the ggnet (Marbach et al., 2016) and network (Butts, 2008) packages.

4. Assign Comment Labels to Topic Bins

We construct a new automation approach to assign comments to their respective topic. After utilizing the aforementioned methods to build an initial topic and keywords table, we conduct an iterative procedure of using the top CLS labels and bin them appropriately. After iterating through the top CLS labels, we handle any comment labels that are left in the *Other* category by referencing the second highest CLS labels and then the third highest CLS labels subsequently after. Through this iterative process, we also alter the topic and keywords table if the need arises. This method facilitates the progression of filtering comment labels into topics bins that may have been categorized as *Other* through previous methods. The details of this process are outlined in Chapter IV.

5. Web-Based Application

As part of our methodology, we construct an R Shiny (Chang et al., 2018) application that serves to illustrate the process and assist the user in their analysis of survey text through graphical visualizations and a user-friendly interface. While the application is still in development, it provides the framework for producing data tables and plots for meaningful analysis. Chapter IV delves into the details of our process and the figures shown in Chapter IV are pulled directly from our application.

III. SURVEY RESULTS

This section covers the background information of the Navy Retention Survey (Navy Standard Integrated Personnel System, CVSS, n.d.) data that we use for our analysis. In 2013, the Career Viewpoint concept was introduced "to improve survey administration in the Navy…[which]… developed into the construction of the Career Viewpoint Surveys and Studies (CVSS) application" (Cairoli, 2017). In 2014, the Bureau of Naval Personnel (BUPERS), Military Community Management (BUPERS-3), with input and flag level endorsement from the Office of the Chief of Naval Operations (OPNAV) Military Personnel Plans and Policy Division (N13), created the Navy Retention Survey.

A. CAREER VIEWPOINT AND THE NAVY RETENTION SURVEY

Career Viewpoint Surveys and Studies (CVSS) is an online application and survey tool that relies on the Navy Standard Integrated Personnel System (NSIPS). It is used to identify sailors through demographics and military related aspects that are available within their Electronic Service Record (ESR). CVSS was "developed with the intention of disseminating the Navy Retention Survey, but the system proves beneficial for surveys that require short turnaround times and increased participation" (Cairoli, 2017). The Navy Retention Survey consists of two approved versions: The Exit Survey (Active Component (AC)/Full Time Support [FTS]) and the Milestone Survey (AC/FTS).

1. Exit Survey

The Exit Survey's (AC/FTS) main purpose is to identify the reasons why sailors choose to leave active duty service. Participants are contacted approximately half a year prior to their Estimated Date of Loss to the Navy (EDLN), a status indicating termination of naval service in Career Waypoints (C-WAY), or a Transition Assistance Program (TAP) eForm indicating that they are leaving with an estimated date of loss. Sailors who have an indication in their record of pending departure from active service are contacted via their official email address in NSIPS. Officers receive the request if they have an EDLN or have orders to exit active service. Members can also utilize their ESR and self-request access,

or a member's career counselor can request the survey through the NSIPS Career Information Management System (CIMS) account.

2. Milestone Survey

This Milestone Survey (AC/FTS) targets sailors with remaining active duty obligation who reside within the time frame of making the decision to stay in or leave active duty. This version is designed to show indications of a member's intent to either remain in active service or not. Fifteen months from a minimum service requirement (MSR), or projected rotation date (PRD) when no MSR exists, officers are able to view this survey. This is about three months before service members must either negotiate for a new set of orders involving another rotation or officially communicate their decision to resign. Enlisted members are sent the survey 18 months before their Soft End of Active Obligated Service (SEAOS); this is a little less than half a year before the reenlistment request process initiates in the C-WAY system. The windows set for officers and enlisted were created to ensure that sailors can effectively communicate their concerns and intentions with the Navy before the process begins to request reenlistment or negotiation for orders (Navy Standard Integrated Personnel System, CVSS, n.d.).

3. Survey Distribution

Both versions of the Navy Retention Survey were initially distributed on 01 July 2014 and are automatically sent out on a monthly basis dependent on predefined criteria (Cairoli, 2017). These surveys are sent to selected individuals and are accessible for eight weeks. Both surveys consist of up to 150 questions and are customized by an individual's response to the 15 core questions as well as their demographics detailed in NSIPS. A majority of these questions use a seven-point scale to represent an individual's intent to stay or leave (Cairoli, 2017). An example is provided in Table 2.

Table 2. Questions Involving Seven Point Scale. Adapted from Cairoli (2017)

On a sliding scale of 1-7, with 7 being the strongest influence to stay, please indicate if the following factors influence you (contribute to your decision) to stay on active duty, leave active duty, or have no effect on your Navy career intentions.	Leave----- 1	2	No Effect 3	4	5	6	-------Stay 7
Promotion/Advancement opportunities	O	O	O	O	O	O	O
Career assignments (number of options, control over PCS assignments)	O	O	O	O	O	O	O
Command climate (previous and current commands)	O	O	O	O	O	O	O
Work-life balance (operational work demand, sea duty, time away from home)	O	O	O	O	O	O	O

In addition, multiple option questions with a comment box to clarify a response and stand-alone comment boxes ranging from 100 to 1000 characters are present within the surveys (Cairoli, 2017).

B. COMMENT ANALYSIS APPLICATION

To demonstrate our methodology as implemented in an R Shiny (Chang et al., 2018) web-based application, we use data collected from the Navy Milestone Survey that asks the respondent to "please list any other factors that have influenced your career intentions." The total response to this field is well over 10,000 comments. Similar to Layug (2018), we utilize 150 randomly selected comments to demonstrate the use of our application. To illustrate the first step in our process of obtaining candidate tokens from a given comment, we begin with a sample comment below.

Example Comment:

"I stay because I have a few years left before I retire and I also would like to continue to put in for a commission to grow in that field of policy and instruction. It's just sad that Enlisted gets treated like serf's and told what they will do and officers get respect on how they will get things accomplished. Everyone is a leader and everyone should be treated as such. \n\nLook at how many of your great sailors over the years have chosen to get out or have been over looked for a MAP after being SOY, EP's etc. Manning and the way we promote should change rather than changing rating NEC's."

1. Preprocess Candidate Labels

Before the comment is partitioned into candidate labels it undergoes a preprocessing phase**.** During preprocessing all comments are converted to lower case and all punctuation besides commas, colons, semi-colons, periods, apostrophes, question marks, and exclamation marks are removed; Furthermore, we replace contractions in a comment with their non-contraction equivalents (Cairoli, 2017). The example comment results in 276 candidate labels. Table 3 illustrates a handful of candidate labels created from the example comment above once it is preprocessed.

Table 3. Example of Preprocessed Candidate Labels

get things accomplished
gets treated like
field of policy
changing rating
for a commission

2. Calculate Variable Values and Computing CLS

Each comment is tokenized into either 1-, 2-, or 3-gram candidate labels; for each candidate label, we calculate their respective variable values utilizing the steps in Chapter II, Section A. Table 4 illustrates a snapshot of candidate labels along with their respective label-specific and comment-specific variables that contribute to the overall CLS. Note that a comment can have multiple candidate labels as illustrated in Table 4. Information pertaining to the definition and details of label specific and comment-specific variables are covered in Chapter II, Section A. After finding every possible candidate label, we extract the candidate label with the highest CLS for a particular comment to represent the comment label. We then use these comment labels that have the highest CLS and conduct several methods of analysis as outlined in the following section to help identify meaningful topics.

Table 4. Candidate Label Variables

ID	Token	Total.Words	Label.Size	Freq	RFO	FH	TT	PTT
doc_id_1	a commission	117	2	1	0.800	1	FALSE	FALSE
doc_id_1	a commission to	117	3	1	0.798	1	FALSE	FALSE
doc_id_1	a few	117	2	1	0.957	1	FALSE	FALSE
doc_id_1	a few years	117	3	1	0.956	1	FALSE	FALSE
doc_id_1	a leader	117	2	1	0.461	0	FALSE	FALSE
doc_id_1	a leader and	117	3	1	0.456	0	FALSE	FALSE
doc_id_1	a map	117	2	1	1.009	0	FALSE	FALSE
doc_id_1	a map after	117	3	1	1.018	0	FALSE	FALSE
doc_id_1	accomplished	117	1	1	0.491	0	FALSE	FALSE
doc_id_1	accomplished everyone	117	2	1	1.009	0	FALSE	FALSE
doc_id_1	accomplished everyone is	117	3	1	1.018	0	FALSE	FALSE
doc_id_1	after being	117	2	1	0.174	0	FALSE	FALSE
doc_id_1	after being soy	117	3	1	1.018	0	FALSE	FALSE
doc_id_1	also	117	1	1	0.879	1	FALSE	FALSE
doc_id_1	also would	117	2	1	0.878	1	FALSE	FALSE

3. Create Topic Bin Key

Creating a topic bin key involves multiple methods as outlined in the following sections. It is through cross referencing these methods that enables the analyst to create their desired topic bin keys.

a. Most Frequent and Salient Keywords List

Our first method of finding potential categories for comment labels relies on frequency. We follow the work of Layug (2018) and create a reference list composed of tokenized comment labels that are broken into bigrams and trigrams. Tables 5 and 6 depict the most frequent bigrams and trigrams from the 150 randomly sampled comments. If there are any new keywords that are considered to be valuable, those keywords are added to the

list of keywords. Note that our process of creating an initial keywords list based on frequency was created after eliminating stop words.

Table 5. Most Frequent Bigrams Reference List

duty station	command morale	leadership climate
job satisfaction	command tour	leadership communication
9 11	constant rollercoaster	leadership mentorship
duty station	good leaders	life circumstances
additional opportunities	core values	low military
additional training	current command	manning review
amazing leadership	dollar equivalent	mcpo macm
assistance friends	educational opportunities	mentorship finances
geographic location	factors navy	micro management
biggest issue	family education	nams comms
business model	generation active	naval aviation
career decision	health mental	naval aviator
career path	honour courage	base housing
chief shelby	leadership climate	numerous officers
civilian style	leadership communication	operational duty

From Table 5 we see frequent bigrams such as “duty station,” “educational opportunities,” “geographic location,” and “dollar equivalent” that can hint at potential topic categories for capturing a variety of comments. For instance, “duty station” can tie into geographic location and it may relate to places where individuals prefer or do not prefer to be stationed. The bigram “dollar equivalent” can potentially relate to how individuals draw comparisons of pay between the military and the civilian sector.

When looking at the frequent trigrams list in Table 6, one can see “options for transfer,” “availability of subspecialty,” “options for detailers,” personality of detailers,” as examples. Trigrams help to relay more information that may not be captured by the frequent bigrams. For instance, “availability of subspecialty” may relate to certain areas of focus for a particular job that a sailor performs in the military. The trigram “options for transfer” may relate to information pertaining to options available between different locations, jobs within the navy, or jobs outside of military service.

Table 6. Most Frequent Trigrams List

availability of jobs	duty station assignment
13 year mark	insult to injury
availability of subspecialty	lack of authority
additional training opportunities	leadership mentorship finances
amount of international	location for officers
aspect to junior	lot of programs
benefits after 20	opportunities for upward
careen in navy	options for transfer
civilian style conditions	pending with pers
compensation for subspecialty	personality of detailers
continuous random unfunded	poor command morale
options for detailers	post 9 11

At this stage, the analyst may identify meaningful topics and their associated keywords. A detailed example will be covered in Chapter IV.

b. LDA Saliency Reference List and Relevancy

The next method uses an LDA model to help identify potential topics. Similar to Cairoli (2017), we create a corpus composed of all the comment labels that will be used for LDA modeling; The LDA model we create helps us formulate potential topics through the evaluation of saliency, frequency, and relevancy of keywords as defined and discussed in Chapter II.

As part of the second method, we look at the most salient words that appear within the comments with saliency defined in Equation (2.3). From Table 7, we see that a handful of words that are the most salient in our example are unigrams like “career,” “command,” “collateral,” “educational,” “family,” and bigrams such as “career path,” “base housing,” and “good leaders.” Tables 5 and 6 contain bigrams and trigrams that describe career or job types and opportunities as well as education and family. The most salient words give a helpful indication that can aid in determining different subjects or topics that we can use to bin our survey comments. The examples pertaining to both of the most frequent and most salient keywords can potentially steer us on an initial path to finding meaningful topics, but one ought not to rely on this first method alone and thus we provide other methods to cross reference our findings.

Table 7. Most Salient Keywords Reference List

career	additional	role
leadership	good	job
navy	current	sea
duty	life	quality
degree	reason	stay
years	training	major
opportunities	way	circumstances
active	benefits	family
command	career path	educational
lack	civilian	lot
active duty	collateral	retirement
amount	doctorate	opportunity
influence	dual	enjoy
problem	good leaders	assistance
retention	gs engineering	base housing
retention problem	huge	benefit
year		bitter

The package LDAvis was developed by Sievert and Shirley (2015) to identify topics within text data. LDAvis, uses R (R Core Team, 2018) to form an interactive visualization for identifying topics using LDA that is internet based.

Depicted in Figure 2, we see the visualization produced using LDAvis shows seven topics that can be referenced when determining our topic bins. The seven topics are depicted as circles on the left of Figure 2. When the analyst selects a topic, the most relevant words for that topic are listed on the right. We note that the analyst has control over the relevancy parameter λ via the slider bar in the top right of Figure 2. Here λ = 0.1. For example, we see that the top 30 most relevant terms of topic 3 (listed on the right of Figure 2) are “shore,” “sea,” “location,” “geographic” near the top and “pcs,” “san,” and “duty” are in the middle. These can be words that have a shared topic in relation to area or location of where one is stationed. Relevant words can also be a possible reference to the jobs in certain locations that people do a permanent change of station (PCS) move to. For instance, “san” is relevant because there are a significant number of comments that refer to San Diego.

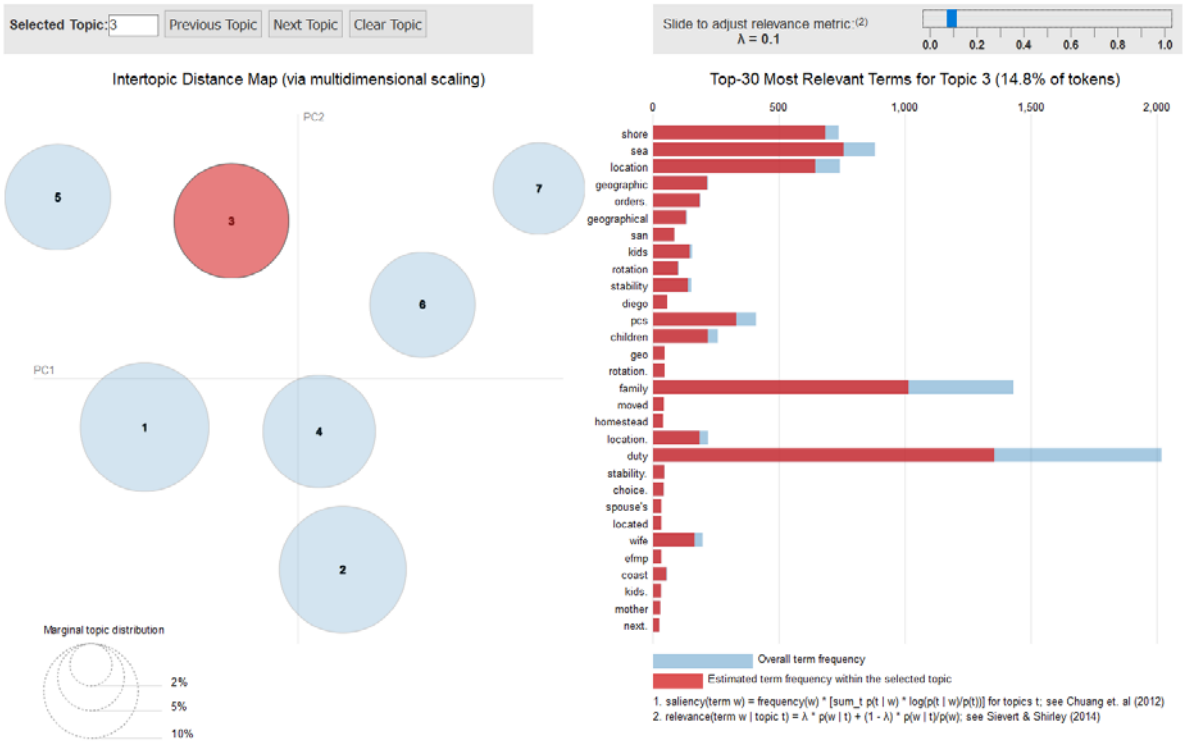

Figure 2. LDA Topic 3 Visualization

Another example, is topic 5 depicted in Figure 3. In this topic, the top relevant terms are “retire,” “retiring,” “20,” “years,” “hyt,” etc. These words are related to time in service completed in the Navy. For instance, “20 years” is the minimum amount of years required in active duty service before a sailor is eligible for receiving full military retirement benefits. Furthermore, “hyt” is an acronym that stands for high year tenure, which is a program with set policy rules and regulations that the Navy uses as an aid to determining advancement and retention of enlisted sailors. There are also entries such as “staying,” “eligible,” “30,” and “19” that may hint that this topic is related to the duration one stays in the military.

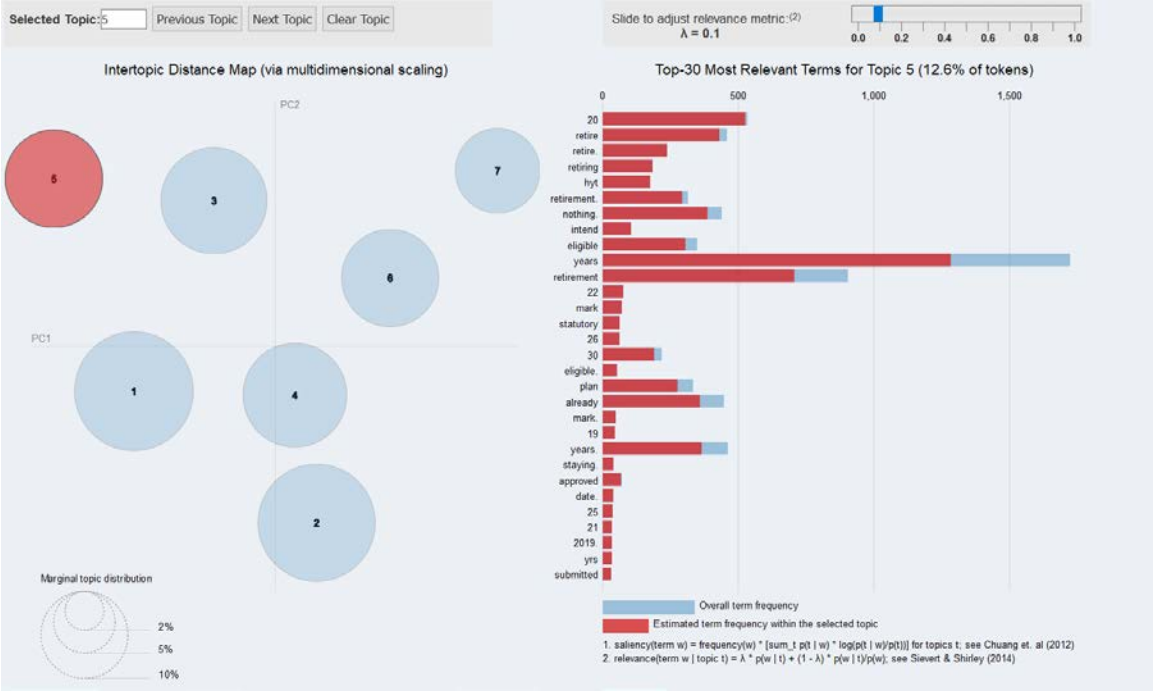

Figure 3. LDA Topic 5 Visualization

c. Correlation Network Method

The next method we use when choosing topic categories is a network method. In this method, we identify the most salient and most frequent 1- to 3-grams and compute the correlations between pairs of terms to formulate a network. Constructing a correlation network allows us to compare unigrams, bigrams, and trigrams as a step in identifying reasonable topic categories that are not already picked up by the other methods (Layug, 2018). Furthermore, creating a correlation network aids in filling in topic bin keys as well (Layug, 2018). Figure 4 illustrates a portion of the correlation network that was formed from our sample data. In this example the bigrams "mentoring aspect," "collateral duties," and "detailers matters" helps the user find possible relationships and potential topic categories not shown in previous methods. For instance, in Figure 4 the "detailers," "matters," "detailers matters," grouping can be cross referenced with Table 6, which contains the "options for detailers" trigram listed as part of the most frequent trigrams list. Cross-referencing between the methods enables the user to establish reasonable and credible topics for categorization.

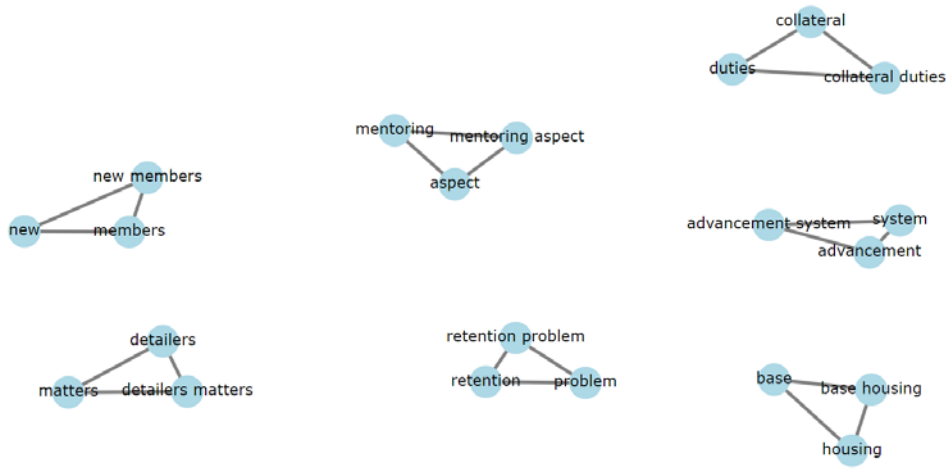

Figure 4. Sample Correlation Network

4. Assign Comment Labels to Bins

By using the methods of the previous section, and through trial and error of our new automation process of assigning comments to topic bins, we refine the number of meaningful topics for binning our sample comments to ten topics. These ten topics are: *Career Opportunities*, *Time*, *Work Environment*, *Family*, *Location*, *Money/Benefits*, *Education*, *Retirement*, *Mental/Spiritual/Physical Health*, and *Navy Policy/Programs*. Table 8 provides an outlook of these main topics and a handful of corresponding comment labels that are pertinent to these topics.

Table 8. Summary of Naval Retention Survey Sample Comments

Topic	Example Labels
Career Opportunities	career decision, availability of jobs, job satisfaction, continuous career, job options, future option
Time	20 years, work hours, long time, time of service, year, work hours
Work Environment	leadership mentorship, pyramid organization, leadership communication, subordinates, combat leaders, micro manage
Family	family, wife, husband, family education
Location	location instruction, other countries, duty station, geographic location, base housing, sea duty
Money/Benefits	dollar equivalent, pay, great pay, benefits
Education	tuition assistance, educational opportunities
Retirement	retire, retirement
Mental/Spiritual/Physical Health	health, medicine, doctor, religion, god, blessing, life, passion
Navy Policy/Programs	nams comms, mcpo macm, 1306 pending, experience in nsw, core values, programs, aviation, choice in orders

Figure 5 represents the topic distribution proportions of the 150 randomly sampled comments pulled from the Navy Milestone Survey that asks the respondent to "please list any other factors that have influenced your career intentions." Over 25% of the comments in this sample are categorized under topic 1 (*Career Opportunities*). The next most frequent topic is topic 3 (*Work Environment*). The third most frequent topic is topic 2 (*Time*) with just over 11%. Topic 8 (*Retirement*) and topic 6 (*Money/Benefits*) both represent less than 5% of the 150 comments. Section B of Chapter IV will go into further detail of how our process is able to provide a quantitative measure of these survey responses and categorize each individual response into their respective topic bins.

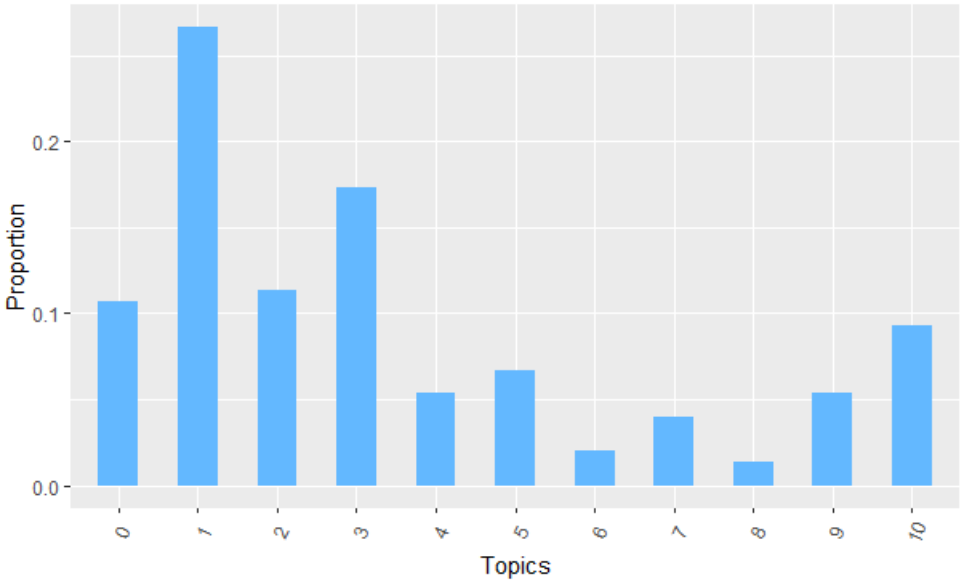

Figure 5. Topic Distribution of Sample Comments

THIS PAGE INTENTIONALLY LEFT BLANK

IV. DISCUSSION AND VALIDATION

A. STEMMING VERSUS LEMMATIZATION

One of the main differences that we implement into the follow-on work of Cairoli (2017) and Layug (2018) is the use of lemmatization in identifying the top scoring candidate tokens. Stemming truncates words into their respective word stem. It utilizes a heuristic approach to determine these particular word stems and differentiates these stems from structural roots that are found within a dictionary. Another way to view stemming is that it breaks a word down to a form that is about the same length or smaller depending on heuristic conditions. Lemmatization, on the other hand, generally refers to handling "things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma" (Manning et al., 2008).

Figure 6 represents a portion of our correlation plots based on the 150 stemmed comment labels. As one can see, the effects of stemming has led to the truncation of important keywords that assist our process in creating topics. In Figure 6, examples such as "advance," "simple," "navi," "retir," "engine," and "collater" are the results of stemming candidate tokens. This can potentially make it challenging for the user since there can be more room for misinterpretation of keywords. For instance, if a sample comment contains acronyms, then there can be some confusion that arises with stemming. There are many acronyms in the Navy and an example is "PERS" which is used in organizing the different departments within Naval Personnel Command (NPC) and can be interpreted in many different ways. In our example above, "per" appears and can cause some confusion in that one may not be sure if it represents "PERS" or was truncated due to stemming and represents "person," "persons," or "personnel."

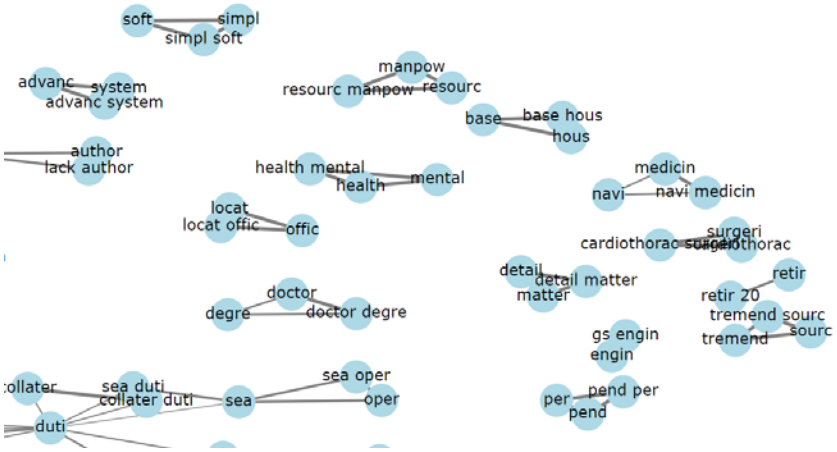

Figure 6. Correlation Network of Stemmed Keywords

Figure 7 represents the correlation plots based on the same 150 comments when using lemmatization. The results appear to be more clear in that there are no partial words displayed when compared to Figure 6. In this sense, the keywords are not as difficult to interpret. In Figure 7, the words illustrated such as “housing,” “mentoring,” “assistance,” and “influencer” are either the words directly from the text in the survey or the canonical forms of the response text found within the survey. In Figure 7, it is easier to determine that “pers” is entered by individuals in response to this question and that it is not a product of the lemmatization process.

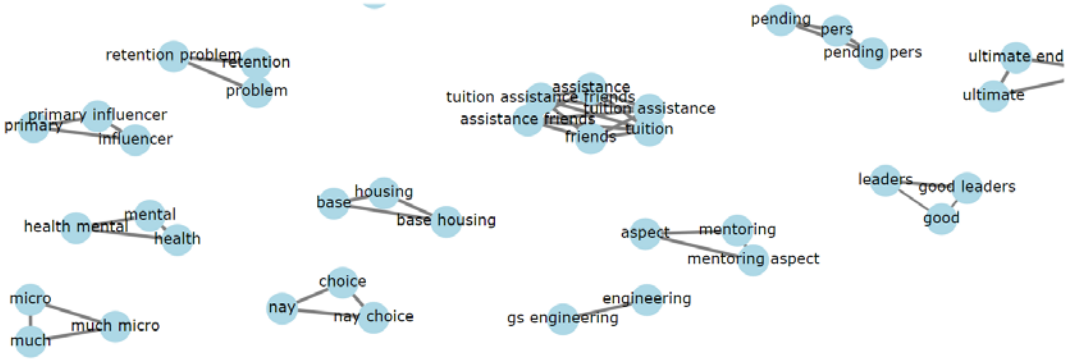

Figure 7. Correlation Network of Lemmatized Keywords

B. LDA RELEVANCY METRIC

Another important implementation that is utilized in our research is the relevancy metric, Equation (2.3), and the LDAvis package (Sievert & Shirley, 2015). The example that is shown in Figure 8 is pulled from the same comment labels from the Navy Milestone Survey data that is used in Chapter III for Figures 2 and 3. In Figure 8, Topic 6 will be used for a relevancy comparison and in this graphic we see that the relevancy parameter λ is set to 1. A few keywords that appear in the top are "pay," "civilian," and "medical." In the middle of Figure 8 (right side), keywords such as "base," "housing," and "bah" appear while "members," "cost," and "bonus" appear at the bottom. When looking at the estimated frequency of the term in the topic depicted in red and comparing it to the overall term frequency depicted in blue, we see the words that stand out in terms of relevancy.

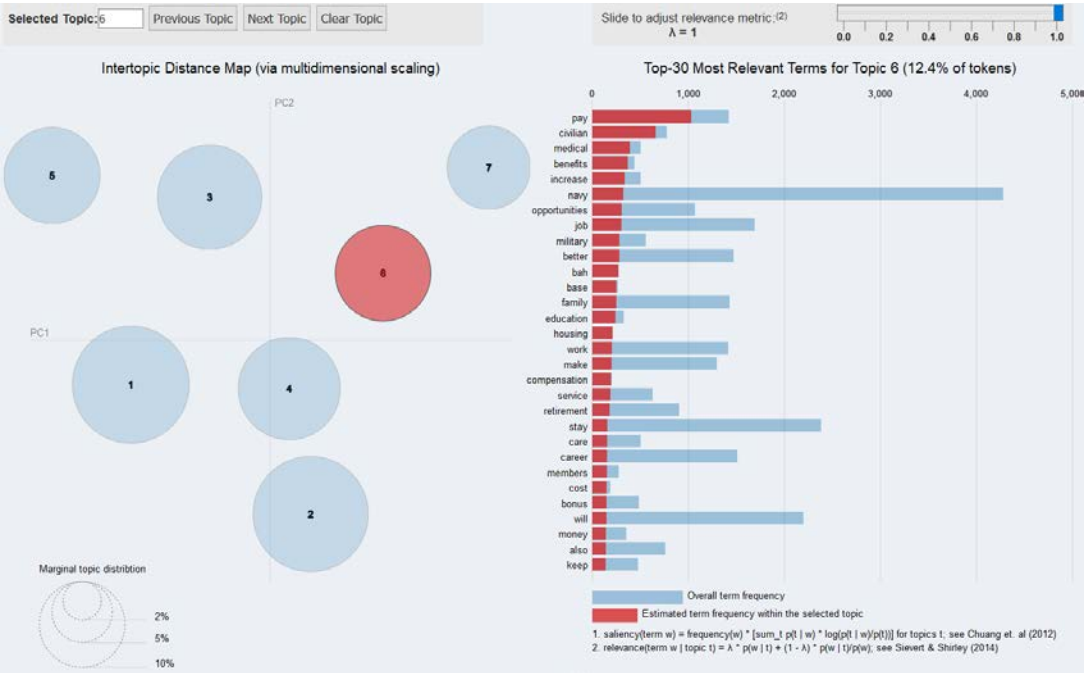

Figure 8. LDAvis Topic 6 Relevancy λ = 1

Now, when looking at topic 6 with the relevancy metric λ set to 0.3, we observe some differences in the keywords that are presented. Figure 9 illustrates the changes that occur to the list of keywords with λ set to 0.3. The terms that appear in the top are "pay," "civilian," and "bah." In the middle (right side), keywords such as "dental," "education,"

and “costs” appear while “degree,” “benefits,” and “allowance” appear at the bottom. There are a few words that have been shifted around such as “bah,” “medical,” and “benefits” as well. When glancing at the estimated topic term frequency and overall term frequency bar chart comparisons, these words are closer in comparison than the keywords that have been listed in Figure 8. These relevant terms appear to have a more localized relationship to topic 6 and can help suggest to the researcher the type of keywords that are closely related as depicted in Figure 9. Overall LDAvis (Sievert & Shirley, 2015) is a useful cross- referencing tool when developing topic bins.

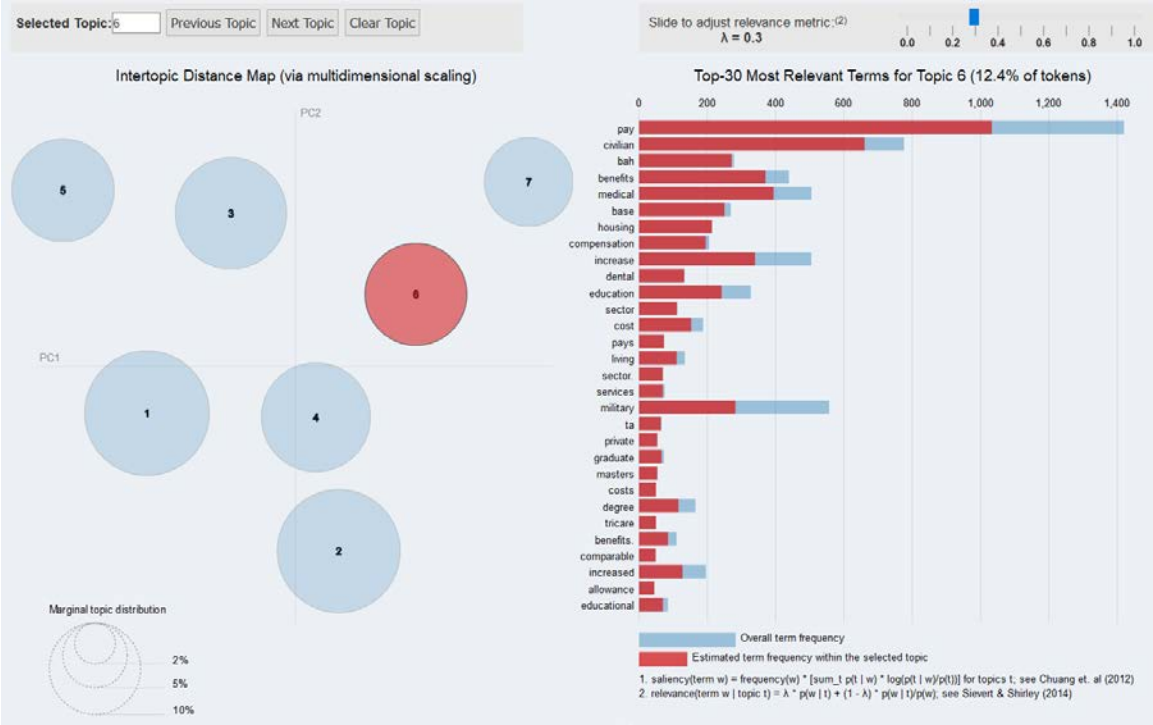

Figure 9. LDAvis Topic 6 Relevancy λ = 0.3

C. ASSIGNING COMMENT LABELS TO TOPIC BINS

After applying the methods covered in Chapter IV Section B, we obtain seven initial topics that are used in initial binning of the comment labels. The seven topics are**:** (1) *Career,* (2) *Time,* (3) *Work Environment,* (4) *Family,* (5) *Location,* (6) *Money, and* (7) *Education***.** Table 9 illustrates these topics along with the relevant keywords that belong to each topic. Figure 10 illustrates the proportion of comment labels that are pertinent to

each of the seven topics. From the initial set of topics, it appears that there is a significant proportion of comment labels that have been left in the (0) *Other* bin.

Cairoli (2017) uses a word cloud of the most frequent and most salient words of comment labels found in the *Other* category to aid the analyst in identifying additional topics. From Figure 10, it is apparent that well over 60 comment labels fall into the *Other* (0) category. However, we will not implement the use of word clouds to handle this predicament and instead apply a new method.

Table 9. Initial Topics Data Table

Topic	Label
Career	screen into more, job, training, eval system, career, job satisfaction, great employer, career decision, command, certain job, availability of jobs, job satisfaction, talent management
Time	20 years, work hours, 20 years, countless times, 20 years, long time, time, time of service, time
Work Environment	admininistrative, physical injuries, previous commands, work hours, few sailors, leadership mentorship, manning review, pyramid organization, good leadership, wise man, eval system, leadership communication, chief, poor command, amazing leadership, my command, options for transfer, good assignments, subordinates, combat leaders, naval aviator, pyramid organization, talent management, personality of detailers
Family	family, wife, family benefits, husband, family education
Location	location instruction, other countries, duty station, geographic location, member on okinawa
Money	dollar equivalent, pay, great pay, pay, retirement benefits
Education	intelligence, additional training, training, tuition assistance, educational opportunities, family education, standstill with schooling

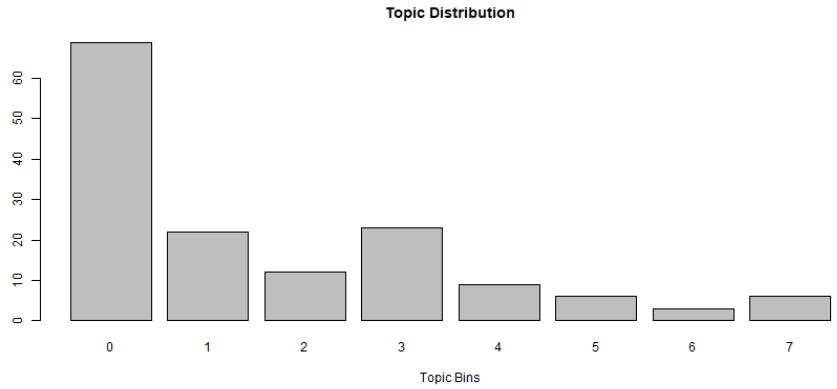

Figure 10. Original Topic Distribution

To re-assign comments in the *Other* category, we utilize their second- and third-best CLS labels. The second- and third-best CLS labels may give insight into what the respective comment is about. They can either be more obscure than the top CLS label, quite similar to the top CLS label, or they can potentially be more descriptive. For instance, in Figure 11, the labels "factors navy," "deserve," "closest i,", and "op t" are the highest CLS labels (as shown from the "List 1" tab in Figure 11) for four separate comments; interpreting these comment labels can be difficult when trying to discern the appropriate topic these comment labels fall into, thus one may feel that the *Other* category is the most appropriate to use.

Figure 11. List 1 Input

However, when looking at Figure 12 we see that the second highest CLS equivalent labels are "base housing," "s do not," "continuous career," and "lack of opportunity." It appears that most of the second highest CLS labels are a bit more informative, which helps when categorizing these comments. Furthermore, being able to look at the next highest CLS labels gives the user more information about how the survey comments can be labeled. This new information allows us to reassess the topics created and formulate new potential topics that are less broad than the initial seven topics. The topics formed are now: (0) *Other*, (1) *Career Opportunities*, (2) *Time*, (3) *Work Environment*, (4) *Family*, (5) *Location*, (6) *Money/Benefits*, (7) *Education*, (8) *Retirement*, (9) *Mental/Spiritual/ Physical Health, and* (10) *Navy Policy/Programs*. Figure 13 illustrates the new distribution

of these 11 topics; as one can see, there is a significant drop from over 65 comment labels to about 35 comment labels binned as *Other*.

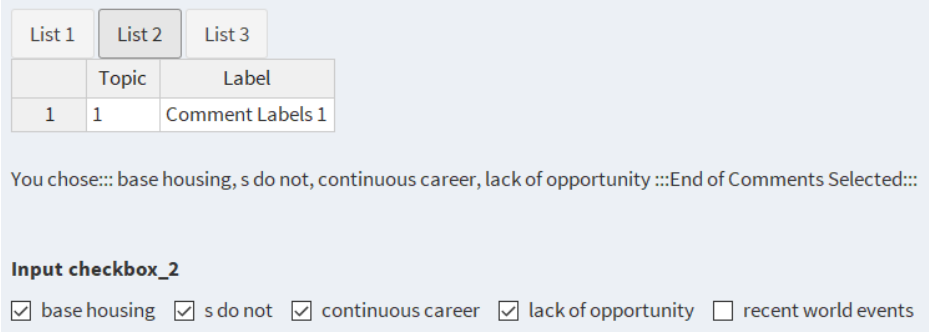

Figure 12. List 2 Input

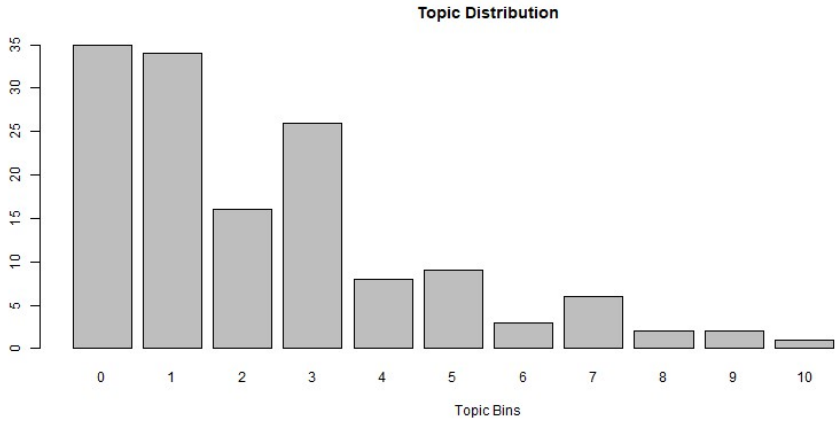

Figure 13. Topic Distribution Using 2rd Highest CLS for *Other*

Although there is some progress with utilizing the second highest CLS to manage the comment labels left in *Other*, there is still a significant number left over. We continue to the third highest CLS labels and analyze the label comments to see if they can provide more information. After the process is complete, we plot the results in Figure 14. Here, we see that there is some progress in dealing with the *Other* comment labels; instead of 35 comment labels there are about 15 comment labels binned in *Other*.

For this example, we will stop at the third highest CLS labels, but the process can continue to the next highest CLS labels when handling comment labels binned in the *Other*

category. When looking at Figure 14, the top 3 topics that contain a significant proportion of the comment labels are (1) *Career Opportunities* with 27%, (3) *Work Environment* with 17%, and (2) *Time* with 11%. For this example, the (6) *Money/Benefits* and (8) *Retirement* both contain less than 5% of the total amount of comment labels.

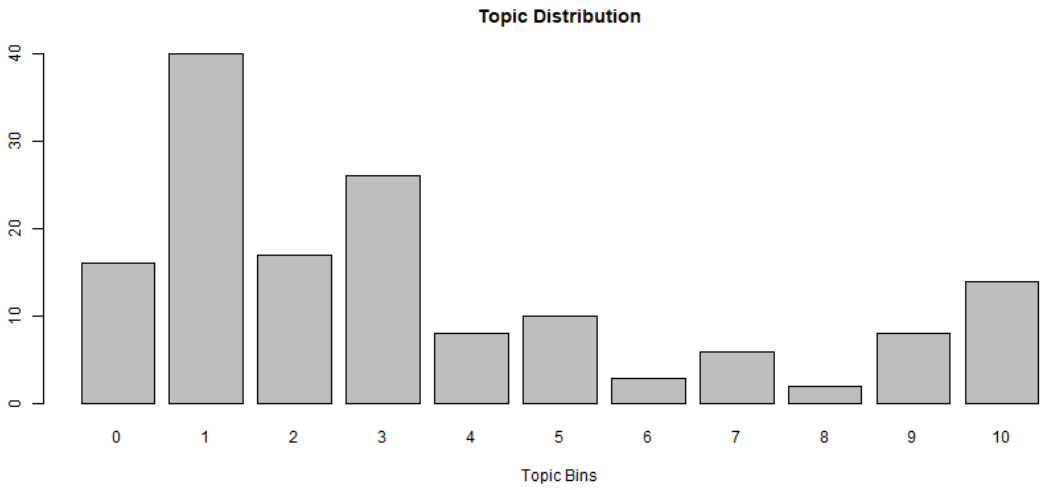

Figure 14. Topic Distribution Using Third-Highest CLS for *Other*

If there are no pre-determined topics that the user has in mind, then this would be the process taken to identify meaningful topics within a given data set. However, if there are already pre-determined topics that the user is given, then the user can add them into the application and bin the comment labels according to their respective topics. The R Shiny (Chang et al., 2018) application we develop is used to categorize the original comments into their selected topics.

D. COMMENT BINNING VALIDATION

1. Expert Binning

The validation process of our new methods involves the same data Cairoli (2017) uses in her research. This is a separate data set than the one used in the previous section. The responses focus on reasons why military personnel leave active duty service and what kind of incentives would make them stay. Two subject matter experts working on the data

at the time conducted separate analysis and provided their binning recommendations. Both subject matter experts were asked to analyze 200 randomly selected comments and bin them to an appropriate topic; There were up to three topics formed in the summary binning process and each expert ranked their selection if more than one topic was provided (Cairoli, 2017). If there were any comments that were not relevant to the provided topics, the experts were then asked to create a new additional topic bin.

2. Comparison

We use 18 bins formed by Cairoli (2017) and the subject matter experts as our topic bins in this validation process. Based on 200 randomly selected responses, we apply our methods to assign comments to topic bins and compare the results to that of Cairoli (2017). About 49.2% of the binned comments for reasons leaving were exact matches to the experts' results. When looking at the binned comments for reasons to stay, approximately 34.1% of the comments were exact matches. Although the percentage of exact matches for the comments binned by our methodology are comparable to those of Cairoli's (2017), our comparisons with the topic bins from the subject matter experts are more auspicious.

Of the 200 randomly selected responses, only the responses that both experts binned are used for comparing our method's results. For Cairoli's (2017) methodology on analyzing "reason leaving" comments, "30.4% of binned comments are exact matches to both experts, 38.1% match the top bin of at least one expert, and 64.9% match one of the top three bins by either of the experts" (Cairoli, 2017). However, Cairoli (2017) notes that both experts agreed on their first-choice bin for about 46.2% of the responses. As for our methodology, approximately 45.1% of the expert binned comments are exact matches to ours, 56.6% of our bins match the top bin of at least one expert, and 74.6% of our bins match one of the top three bins by either of the experts.

As for the stay results, Cairoli's (2017) methodology shows that "27.3% of our binned comments are exact matches to both experts, 45.5% match the top bin of at least one expert, and 56.8% match one of the top three responses by any of the experts with 43.2% of the primary ranked topic matching for the experts" (Cairoli, 2017). When implementing our methodology, our results show that 25.0% of our binned comments

are exact matches to both experts, 40.9% match the top bin of at least one expert, and 68.2% match one of the top three responses by any of the experts.

As Cairoli (2017) points out, for a random sample of 200 comments, even two Navy manpower subject matter experts categorized nearly half of the comments into different bins. Through Cairoli's (2017) methodology "approximately 65% of at least one identified bin when bins are provided is comparable to levels attained by Chuang et al. (2012b) when labeling larger documents" (Cairoli, 2017). When looking at our results for the "reason leaving" comments, about 75% of our bins match one of the top three bins identified by either of the experts and about 68% of our bins match one of the top three bins identified by either of the experts for the "reason for staying" comments.

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

Our adaptation of the approaches to natural language processing developed by both Cairoli (2017) and Layug (2018) seeks to provide a deeper understanding of the free response questions posed in Navy surveys. Open response surveys can provide invaluable insight that can potentially alter or promulgate a variety of policies. Through cross referencing multiple methods for finding meaningful topics, our new analytical approach involving an R Shiny (Chang et al., 2018) application further enables researchers to build intuition in their analysis of free response text. With the amount of data available and with limited time for analysis, our methodology can enable researchers to identify and organize free response text with relative ease. This can potentially allow analysts to discern important information and that can help weigh and possibly determine which courses of action to take.

Although our methodology is demonstrated on Navy Milestone Survey and Navy Exit Survey data, it can apply to a variety of surveys distributed throughout the Navy. As mentioned by Cairoli (2017), there are well over 65 surveys that include short response text conducted by the Navy annually and many survey analysts are open to exploring tools to aid their research endeavors. Furthermore, our new automation process of binning topics also addresses the previous challenge of handling the leftover comment labels that are set into the *Other* category. By utilizing the next highest CLS labels and allowing the user to bin them according to a relevant topic, one can readily handle a significant amount of obscure comment labels that would have been previously left as *Other*. With our refined approach in processing short-text comments, Navy survey analysts can potentially discover prevalent topics within a timely and decisive manner.

B. FUTURE WORK

Although our work expands on the foundations laid by Cairoli (2017) and Layug (2018) for identifying meaningful topics extracted from short-response text, there is more work that can be further improved upon. There are a variety of methods that can be

implemented in future work that can provide more insight into the natural language processing of survey comments.

1. Identifying the Sentiment of a Comment

Sentiment analysis, which deals with extracting subjective information from text, is a field that is yet to be explored in our methodology. Many surveys include open ended response questions, which promotes comments that result in the responder writing with a "positive or negative," "like or dislike" aspect. Determining if a given comment is positive or negative in relation to a particular topic is possible through sentiment analysis. An approach that delves into text mining the sentiment of survey responses can underline pertinent information not previously discovered through our current methods in this thesis.

2. Refining R Shiny Application

Our methodology implements the development of an interactive application involving the R Shiny (Chang et al., 2018) package. Our research provides a new framework for automating the binning procedure of comment labels into their respective topics and provides graphical tools to visualize the topic distribution of comment labels. Our application illustrates the methods and procedures we take when processing and analyzing survey data. However, there is still room for improvement. The user interface of our application can be further enhanced to allow the user to look at a sentiment analysis portion of survey responses, which has yet to be done with our research. In relation to Layug's (2018) work, the application can potentially be fit to process a new reference corpus or upload an existing reference corpus that can be used when calculating the CLS. Furthermore, it is possible to publish and distribute our R Shiny (Chang et al., 2018) application so it can be more easily accessible to several researchers in the future.

LIST OF REFERENCES

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

Butts, C. T. (2008). "Network: A package for managing relational data in R." *Journal of Statistical Software*, *24*(2). Retrieved from http://www.jstatsoft.org/v24/i02/

Cairoli, C. M. (2017). *Categorization of survey text utilizing natural language processing and demographic filtering.* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle.net/10945/56109

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). shiny: Web application framework for R. Retrieved from https://CRAN.R-project.org/package=shiny

Chuang, J. (2013). *Designing visual text analysis methods to support sensemaking and modeling* [Doctoral dissertation, Stanford University]. https://nlp.stanford.edu/~manning/dissertations/Chuang-Jason-dissertation.pdf

Chuang, J., Manning, C., & Heer, J. (2012a). *Termite: visualization techniques for assessing textual topic models*, ACM. doi:10.1145/2254556.2254572

Chuang, J., Manning, C., & Heer, J. (2012b). Without the clutter of unimportant words. *ACM Transactions on Computer-Human Interaction (TOCHI), 19*(3), 1–29. doi:10.1145/2362364.2362367

Kullback, S., & Leibler, R. (1951). On the information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved fromhttp://www.jstor.org/stable/2236703

Layug, C. (2018). *Extracting major topics from survey text responses using natural language processing* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle/10945/60425

Manning, C., Raghavan, P, & Schütze, H. (2008). *Introduction to information retrieval*, Cambridge University Press.

Marbach, M., & Briatte. F (2016). Ggnet2: Retrieved from http://github.com/briatte/ggnet

Navy Standard Integrated Personnel System, CVSS. (n.d.). Career viewpoint retention survey. Retrieved from https://nsipsprod.nmci.navy.mil/

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Sievert, C., & Shirley, K. (2015). LDAvis: Interactive visualization of topic models. Retrieved from https://CRAN.R-project.org/package=LDAvis

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly.

Wijffels, J. (2018). udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP'. Toolkit. Retrieved from https://CRAN.R-project.org/package=udpipe

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California