NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

EXTRACTING MAJOR TOPICS FROM SURVEY TEXT RESPONSES USING NATURAL LANGUAGE PROCESSING

by

Christine Layug

September 2018

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

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

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

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE EXTRACTING MAJOR TOPICS FROM SURVEY TEXT RESPONSES USING NATURAL LANGUAGE PROCESSING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Christine Layug	
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)

In this thesis, we enhance the comment analysis approach of Cairoli's 2017 Naval Postgraduate School thesis, to help the fleet analyze comment responses in Department of Defense surveys and utilize the results to make important decisions. This methodology automates applying descriptive labels to a comment and then uses those labels to categorize comments into a small set of meaningful prevalent topics. We apply this methodology to comments from two recent surveys: a command climate survey as well as an investigation survey looking into the recent increase of physiological episodes experienced by T-45 and F/A-18 aircrews. When applying novel approaches to different data, unexpected matters emerge. These matters shed light on areas of the approach that may need expansion or modification. Motivated by our analysis of text comments from two very different Navy surveys, we extend Cairoli's approach in four ways. Our modifications lead to a generalized model; an approach independent of the need to acquire and preprocess an external reference corpus; more automation of the topic discovery process; and an added element that allows a comment to have more than one topic.

14. SUBJECT TERMS text analysis, SURVEY, Naval Inspector General , Defense Equal Opportunity Management Institute, DEOMI, Organizational Climate Survey, DEOCS	**15. NUMBER OF PAGES** 71
	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.

EXTRACTING MAJOR TOPICS FROM SURVEY TEXT RESPONSES USING NATURAL LANGUAGE PROCESSING

Christine Layug
Lieutenant, United States Navy
BS, U.S. Naval Academy, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2018

Approved by: Lyn R. Whitaker
Advisor

Christine Cairoli, OPNAV N1T
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In this thesis, we enhance the comment analysis approach of Cairoli's 2017 Naval Postgraduate School thesis, to help the fleet analyze comment responses in Department of Defense surveys and utilize the results to make important decisions. This methodology automates applying descriptive labels to a comment and then uses those labels to categorize comments into a small set of meaningful prevalent topics. We apply this methodology to comments from two recent surveys: a command climate survey as well as an investigation survey looking into the recent increase of physiological episodes experienced by T-45 and F/A-18 aircrews. When applying novel approaches to different data, unexpected matters emerge. These matters shed light on areas of the approach that may need expansion or modification. Motivated by our analysis of text comments from two very different Navy surveys, we extend Cairoli's approach in four ways. Our modifications lead to a generalized model; an approach independent of the need to acquire and preprocess an external reference corpus; more automation of the topic discovery process; and an added element that allows a comment to have more than one topic.

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

AFO	absolute first occurrence
App	application
CO	Commanding Officer
CTS	candidate token score
DEOCS	DEOMI Organizational Climate Survey
DEOMI	Defense Equal Opportunity Management Institute
DoD	Department of Defense
DoN	Department of the Navy
DTM	document term matrix
FH	first half
Freq	frequency
LDA	Latent Dirichlet Allocation
MWR	Morale, Welfare and Recreation
NAVINSGEN	Naval Inspector General
PE	physiological episode
PEAT	Physiological Episode Action Team
PET	Physiological Episode Team
POS	parts of speech
PTT	partial technical term
RC	reference commonness
RFO	relative first occurrence
TS	token size
TT	technical term

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Without the proper analytic tools, manually reading survey comments is tedious, time consuming and susceptible to human error. Agencies throughout the Department of Defense (DoD) recognize the information found within survey comment text is a valuable resource and they continually look for tools to analyze or classify survey comments with ease and accuracy (C. Cairoli, personal communication, March 9, 2018).

In this thesis, we enhance the comment analysis approach of Cairoli (2017) to help the fleet analyze comment responses in DoD surveys and utilize the results to make important decisions. This methodology automates applying descriptive labels to a comment and then uses those labels to categorize comments into a small set of meaningful, prevalent topics. When applying novel approaches to different data, unexpected matters emerge. These matters shed light on areas of the approach that may need expansion or modification. Motivated by our analysis of text comments from two different Navy surveys, we extend Cairoli's (2017) approach in four ways. Our modifications lead to a more generalized model; an approach independent of the need to acquire and preprocess an external reference corpus; more automation of the topic discovery process; and an added element that allows a comment to have more than one topic.

The first step of our approach is to parse each comment into "items." This allows each comment to be associated with more than one topic. Comments are parsed into items by number if the comment is a numbered list. Otherwise, they are parsed into sentences or partial sentences at periods and commas, respectively. Our next step is to construct a set of candidate labels for each item. To do this, we break down each item into tokens, which are consecutive-word phrases or *n*-grams. A Candidate Token Score (CTS) is assigned to each unique candidate token. The CTS is a linear function of statistical and linguistic variables that help describe a token's potential for describing the comment or item. The tokens with the maximum CTS for each item become the labels for the comment. To calculate the CTS, we assign two types of variables to each token, token specific variables and comment specific variables.

The token specific variables are “unique to each candidate token and independent of the comments associated with the token” (Cairoli, 2017). The first variable, token size (TS) is a categorical variable indicating the number of words in the token. The second variable, technical term (TT) is a binary variable indicating whether a token is a technical term or not. For text analytic purposes, a technical term, according to Chuang, Manning, and Heer (2012), is a multi-word phrase that meets a specific pattern. Like Cairoli (2017), we adapt the definition of a technical term to follow this pattern: “it begins with either an adjective or noun, strings together adjectives, nouns or prepositions in the middle, and ends in a noun.” The third token specific variable, partial technical term (PTT) is a binary variable indicating whether the token is a substring of a technical term or not.

Comment specific variables are unique to each comment and are factored into the CTS computation. We use the same three comment specific variables as Cairoli (2017): “Freq, the frequency of the token in each comment; RFO [relative first occurrence], a measure of the first occurrence of a token relative to a token of the same frequency; and FH, an indication of whether a token is contained in the first half of [an item] or not.”

We calculate the CTS using estimated regression coefficients similar to the approach of Chuang et al. (2012). Using the essence of their work, we randomly select 180 comments from each of our two surveys that have more than five words. We read each comment, choose a 1- to 3-gram consecutive-word token that best describes the comment and store it as an expert label. In addition, for each comment we randomly select ten tokens (excluding the expert label) to use as false-positives. This produces a data set of 3,960 comments with expert and randomly chosen labels. We also combine this data set with Cairoli’s (2017) data set, which spans three different surveys and contains 2,200 comments with expert and randomly chosen labels. This gives a data set with comments taken from five different Navy surveys. We compute the comment and token specific variable values for each label-comment pair and fit a logistic regression where the response variable is 1 if the label from the comment is an expert label and 0 otherwise to estimate a new set of regression coefficients. The candidate tokens for each item are scored using the estimated coefficients. The candidate token with the maximum CTS among candidate tokens for an

item is assigned as a label for each corresponding item. Thus, each comment may have more than one descriptive label, one per item from that comment.

Once each comment's descriptive labels are identified, we take the analysis a step further by finding meaningful topics among the labels and sorting the labels (and corresponding items) into topic bins. We also provide a tool that automatically constructs a list of potentially meaningful 1- to 3-gram keywords for the item assignment process. Using a systematic approach that evaluates the relationships among the item labels, we create our initial keyword list in three ways. From the labels, the first method extracts the most frequent bigrams and trigrams that do not contain stop words such as "the," "and" or "of." The only exception for allowing stop words is if a stop word appears in the middle of a trigram. The second method evaluates the unigrams and bigrams from item labels using a Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) model and a saliency measure to identify phrases or words that are important, but not overly frequent. The final method uses a network approach to evaluate the correlation between frequent, salient and distinct unigrams from the labels. The resulting 1- to 3-grams from these three methods become keywords that provide a starting point for determining topic bins.

With an initial list of keywords, the analyst can review visual plots and use background knowledge of the survey subject, along with subject matter expertise, to verify that the topic bins make sense and are meaningful. Keywords are compared to the labels and original comment text to group the items into meaningful topic bins. Once assigned to topic bins, text comment responses can be summarized and further analyzed using quantitative methods, such as displaying the topic frequency distribution for a particular question or studying its association with responses to other questions.

The analysis from this methodology not only provides quantifiable results, but gathers quality information more quickly compared to reading all the comments. In our work, we apply the comment analysis methodology to the text comments from two separate surveys, focusing on one question from each survey. The first survey is an organizational assessment administered to a Navy command to gather members' perceptions of their command climate. The second survey is part of an investigation into the recent issues of physiological episodes (PE) within Naval Aviation.

In the organizational assessment survey, we analyze 149 comments to the question "What is one thing your command can do to reduce your stress level?" The top three topics mentioned in these comments are *Job (28%)*, *Leadership (23%)* and *Schedule (20%)*. With this methodology, the command has a more precise understanding of the types of programs, changes or considerations they need to implement to reduce their command's stress levels to improve overall climate and morale.

For the PE survey, this comment analysis approach provides measurable results for the reasons F/A-18 and T-45 aircrews and maintainers believe there was an increase in PE within Naval Aviation. The results of this survey provide information that help focus the investigation into the plausible root cause(s) of PE. The most frequent responses found among the 1,060 comments pertain to specific aircraft parts or equipment and to an increase in awareness of the issue.

References

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

Cairoli, C. M. (2017). *Categorization of survey text utilizing natural language processing and demographic filtering.* (Master's thesis). Retrieved from http://hdl.handle.net/10945/56109

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

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Lyn Whitaker. Thank you for all the encouragement and guidance throughout the development of this thesis. Thank you as well for graciously accommodating my frequent impromptu visits, questions and multiple drafts. Your feedback, constant reassurance and positivity meant so much to me! I am immensely grateful.

A very special acknowledgement and definite credit to my second reader, Christine Cairoli, whom without her diligent work a year prior, this thesis would not have been possible. Thank you, Christine, for sparking my interest in this topic and especially for being so available to answer my questions and for your willingness to help, even during your busy schedule.

I would also like to recognize the assistance of Dr. Larry Shattuck for providing the Physiological Episodes Survey data to fulfill this research. Appreciation should also go to the following NPS classmates for all the help, group study sessions and memorable experiences over the past two years: Rey Cabana, Carlos Cervantes, Brendan Bunn, Charlie Deibler, Jack Li, John Renquist and Pat Saluke.

Thank you to my family for their love and support, as well as for the many visits that were a welcome retreat from the stresses of graduate school. Thank you to my fiancé, Joshua, who could always empathize with me about thesis struggles and provide encouragement and laughs along the way. I am so appreciative of your love, prayers and support that got me through every day at NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Without the proper analytic tools, manually reading survey comments is tedious, time consuming and susceptible to human error. Examining only a subset of the survey comments is inadequate to capture the critical ideas or trends not found in the closed-ended questions of the survey. While commercial text analysis software are available to overcome this limitation, they are expensive, proprietary (Weiss, Indurkhya, & Zhang, 2010) and therefore, impractical for Department of Defense (DoD) surveys. In addition, these general, all-purpose software tools are not tailored to specific language and jargon found in DoD survey comments. Agencies throughout the DoD recognize the information found in survey comment text is a valuable resource and they continually look for tools to analyze or classify survey comments with ease and accuracy (C. Cairoli, personal communication, March 9, 2018).

In this thesis, we enhance the comment analysis approach of Cairoli (2017) to help the fleet efficiently and accurately analyze comment responses in DoD surveys and utilize results to make important decisions. Our method follows that of Cairoli (2017) with modifications inspired by difficulties faced when tailoring the methodology to our specific survey data sets. Each data set consists of "documents" corresponding to text comment responses. The corpus is the set of all documents for a specific question in a particular survey. In our work, we apply a comment analysis approach to the text comments from two separate surveys, focusing on one question from each survey. The first survey is an organizational assessment administered to a Navy command to gather members' perceptions of their command climate. The second survey is part of an investigation into the recent issues of physiological episodes (PE) within Naval Aviation.

A. BACKGROUND

The most common methods for discovering latent topics among a collection of text documents are topic models, which include Latent Dirichlet Allocation (LDA) models (Blei, Ng, & Jordan, 2003). Although most of the successful applications of topic models are for longer documents (Chuang, Manning, & Heer, 2012b), the Office of People

Analytics, the premier analytics organization for the DoD, successfully applied LDA directly to their large-scale data set in DoD surveys to over 24,000 comments (L. Davis., A. Harris, & J. Schneider, presentation, April 3, 2018.) However, for smaller data sets with fewer than 1,000 survey comments, topic models prove less effective. With current survey practices in the fleet moving to a self-service system, any Navy service member can be tasked with analyzing surveys, with the expected number of comments per survey being in the range of only a few hundred to a few thousand (Cairoli, 2017). Further, typical Navy survey comments are either too limited in quantity, are very domain specific or laced with technical terminology that is difficult to understand if not read in context. The recent work of label selection (Chuang et al., 2012b) addresses this second limitation by taking into account linguistic properties. Cairoli (2017) adapts this label selection approach so that it is applicable to survey comments of a smaller scale.

Cairoli (2017) provides a two-step comment analysis approach. The first step uses label selection. Each comment is preprocessed and tokenized into 1- to 3 consecutive-word combinations known as *candidate tokens*. For each candidate token, Cairoli (2017) assigns two types of statistical and linguistic variables, comment specific variables and token specific variables. These variables, such as the token's position in the comment, give an indication of how well the token describes the comment. A separate corpus, the reference corpus, plays an important role in constructing variables that identify tokens associated with technical terms and jargon specific to a survey. The reference corpus can be any document relevant to the specific survey topic and can vary from survey to survey. With the comment specific and token specific variables, a score is then calculated as a linear function of these variables using estimated regression coefficients. The token with the highest score becomes the label for that comment.

The second step of the process uses the descriptive labels, along with more traditional visual text mining methods and LDA, to create primary topic bins. In this step, the analyst uses expert verification for further topic discovery and validation. Using a systematic approach, the analyst sorts the labeled comments into topic bins that correspond to meaningful categories. The categorization of comments into topic bins provides the analyst and decision maker with results that are quantifiable and objective.

Cairoli (2017) applies the comment analysis methodology to the Navy Retention Survey to provide an analysis to the questions "Why are sailors leaving?" and "What will make sailors stay on active duty?" By providing analysts with a means to further filter the topics assigned to each comment by demographic and military related variables, such as rank, community and gender, Cairoli (2017) delivers objective and quantified results that allow retention policy makers to "review, modify and create more relevant incentives to retain our best sailors while working within budget constraints and meeting fiscal year end strength and operational requirements." Cairoli (2017) further validates the methodology on the Female Dress Uniform & Cover Survey administered by OPNAV N1, a survey administered to collect feedback on new or recently modified female uniforms. The methodology by Cairoli (2017) provides an analytic tool to help the Navy analyze comments in a way not previously possible.

B. THESIS OUTCOMES

When applying novel approaches to different data, unexpected matters emerge. These matters shed light on areas of the approach that may need expansion or modification. By analyzing text comments from two very different Navy surveys, we extend Cairoli's (2017) approach in four ways. Our modifications lead to a more generalized model, an approach independent of a reference corpus, more automation of the topic discovery process and an added element of the algorithm that provides multiple labels per comment.

1. Generalized Model

Calculating a score for label selection requires fitting a logistic regression on a hand-labeled portion of data. The expert or analyst constructs this data set by reading a fraction of the comments and selecting a label applicable to each comment. These comment-label pairs are used to estimate a set of regression coefficients. In this research, we combine the comment-label pair data from Cairoli (2017) with similarly constructed data from the two surveys studied in this thesis to estimate generic model coefficients and thus minimize future requirements for analyst hands-on intervention. We demonstrate that these coefficients are generally applicable for use in Navy surveys of different types and allow for even faster prospective comment analysis.

2. Reference Corpus

The approach described in Chapter I, Section A relies on identifying and preprocessing a valid external reference corpus for establishing a basis of the language or technical terminology in a survey. The reference corpus "can be the document that the token comes from, the entire corpus of documents being labeled, or an entirely separate corpus created from general web scraping" (Cairoli, 2017, p. 10). The original approach uses a reference corpus to provide a measure of how common a token is. This measure of reference commonness is then used in scoring the token for label selection. Further, the reference corpus provides a list of technical terms also used in token scoring. While the reference commonness can still be a valid factor for label selection, we use an internal and independent approach that does not rely on an external reference corpus, but focuses solely on the syntactic properties of each word in the context of the comment.

3. Automation of Initial Topic Bin Key

Automating construction of an initial topic bin key delves into an avenue of future work suggested by Cairoli (2017). Chuang et al. (2012a) state, "[real]-world deployments of topic models often require intensive expert verification and model refinement." To minimize expert or analyst intervention and aid replicability, we provide tools that automate constructing an initial list of potentially meaningful topics. The method involves assessing relationships among the comment labels. The words with significant associations are extracted to establish an initial topic bin key list or a plot for visual analysis, as appropriate. The resulting initial topic bin key list and plots offer a starting point for the analyst to review and finalize a topic bin key. These tools help reduce subjectivity and offer an added level of support for faster analysis.

4. Multiple Labels

Motivated by another avenue of future work suggested by Cairoli (2017), this research modifies the initial comment preprocessing to allow multiple labels to be applied to each comment. This modification captures more prevalent topics or ideas in the comments, providing superior feedback without sacrificing computation time.

C. ORGANIZATION OF THESIS

Chapter II explains the methodology. Chapter III provides background on the survey data we use to demonstrate our methodology as well concrete examples of our approach and the surveys' results. The focus of this thesis is not on the results of the surveys, but on improving the tools or process for analyzing comments in DoD surveys. Chapter IV provides a discussion of these improvements. Chapter V concludes with recommendations for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

II. METHODOLOGY

In this chapter, we outline the approach we take to analyze text comment responses in surveys. Our approach follows that of Cairoli (2017) with some important modifications. These modifications are in response to difficulties faced in labeling the text comments for the two surveys described in Chapter III. The description of our methodology is patterned after that of Cairoli (2017). We start with a description of an automated process for extracting candidate labels for each text comment. The second section describes how these candidate labels are then used to assign topics. Although this part of the approach still requires hands-on analysis, we introduce tools for automating portions of the process.

A. CANDIDATE LABELS

The first step of the approach is to construct a set of candidate labels for each comment. Many text analysis methods start by parsing text into individual words, known as "bag of words" analyses. Since that method removes the words from the comment context, bag of words analysis is generally limited to using raw term frequency as a measure for determining labels and fails to take into consideration semantics or other linguistic descriptors. Rather than the bag of words approach, we construct our candidate labels using cautious parsing and preprocessing to filter noise, while considering the linguistic and positional properties as well as the technical terms in a comment. In this research, we modify the initial comment preprocessing to provide multiple labels per comment. We do this by partitioning each comment into candidate items as described in this section, where each candidate item is assigned a label.

1. Candidate Items

For some of the questions in these surveys, respondents format their comments as a list. Thus, a single label is insufficient to describe the entire comment. We illustrate and discuss the benefits of the multiple label algorithm in Chapter IV. In order to capture all the listed ideas in a comment, we first parse the comment into "items." While there are many ways to parse a comment, we take a minimalist approach to make this method as robust as possible. We parse comments in three ways: lists, sentences and partial sentences.

For comments that are in list format, where each item is preceded by a number, we parse corresponding to the list. For our particular data sets, we assume the order in which items appear in a list is not significant and that lists in a comment contain no more than four items. This generally avoids splitting comments unnecessarily after sentences that end in words such as "F/A-18," "T-45" or "2016." Comments that are not in a list format are parsed into sentences and partial sentences by splitting the comment into items at periods and commas, respectively. Splitting at commas captures labels that might appear after words such as "also" and "additionally" or ordinal words such as "first," "second" or "third.

These items become the set of *candidate item*s for a particular comment.

2. Preprocess Candidate Items

We minimally preprocess the candidate items by converting text to lower case and converting common contractions to their whole word equivalent. Similar to Cairoli (2017), we do not remove stop words such as "and" or "the" since the comments are short and these words may add necessary descriptors. We also do not stem or convert words to their root form (Zhai & Massung, 2016). Each candidate item is assigned a unique identification code corresponding to the original comment and the set of candidate items are saved as a corpus of "documents," where each document is the preprocessed text of a candidate item.

3. Candidate Tokens

Each preprocessed candidate item is broken down further into tokens. Tokens are consecutive-word phrases or *n*-grams, which are any *n* consecutive words in an item. Chuang et al. (2012b) demonstrate that there is little added benefit to using more than three words in a label. As a result, we tokenize the candidate items into unigrams, bigrams and trigrams, excluding single stop words. Each item's tokens form a set of candidate tokens.

4. Candidate Token Score

Once each candidate item is broken down into *n*-gram tokens, we evaluate the tokens to determine which one will best represent the item. We assign a Candidate Token Score (CTS) to each unique candidate token. In an item's set of candidate tokens, the token

with the highest CTS becomes the label for that item. The combined labels for the items of a comment then become the set of labels for the comment.

Figure 1 and Figure 2 demonstrate how a listed comment and non-listed comment, respectively, are broken down into their corresponding items and tokens. For visual purposes, we display only the token with the highest CTS. These maximum CTS tokens become the set of labels for the comment.

The CTS is a linear function of statistical and linguistic variables that help describe a token's potential for describing the comment or item. To calculate the CTS, we first assign two types of variables to each candidate token, token specific variables and comment specific variables.

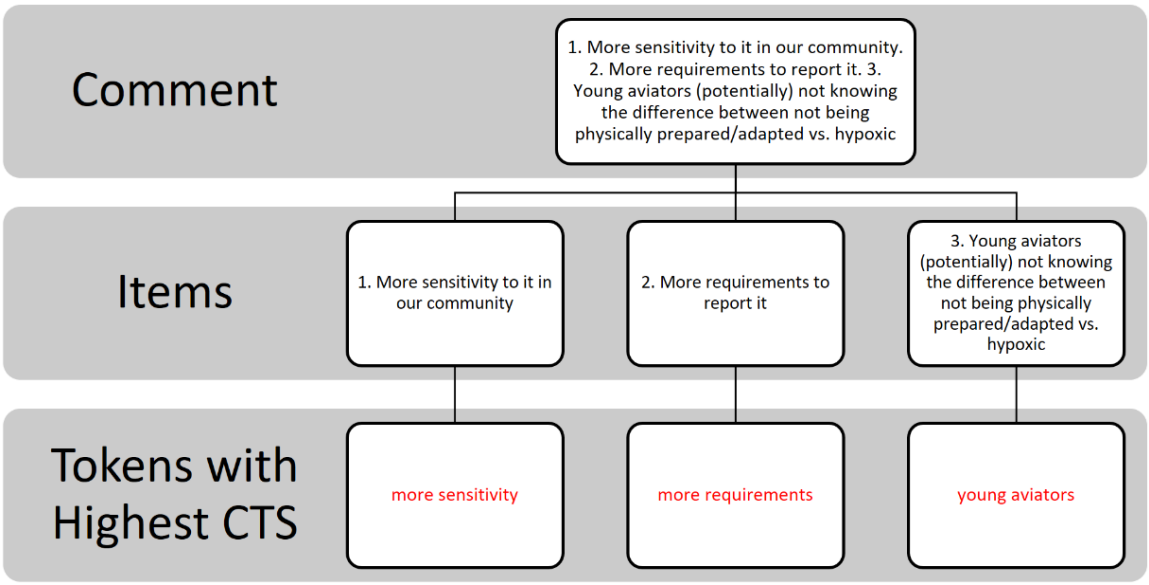

Figure 1. Breakdown of a Comment in a List Format

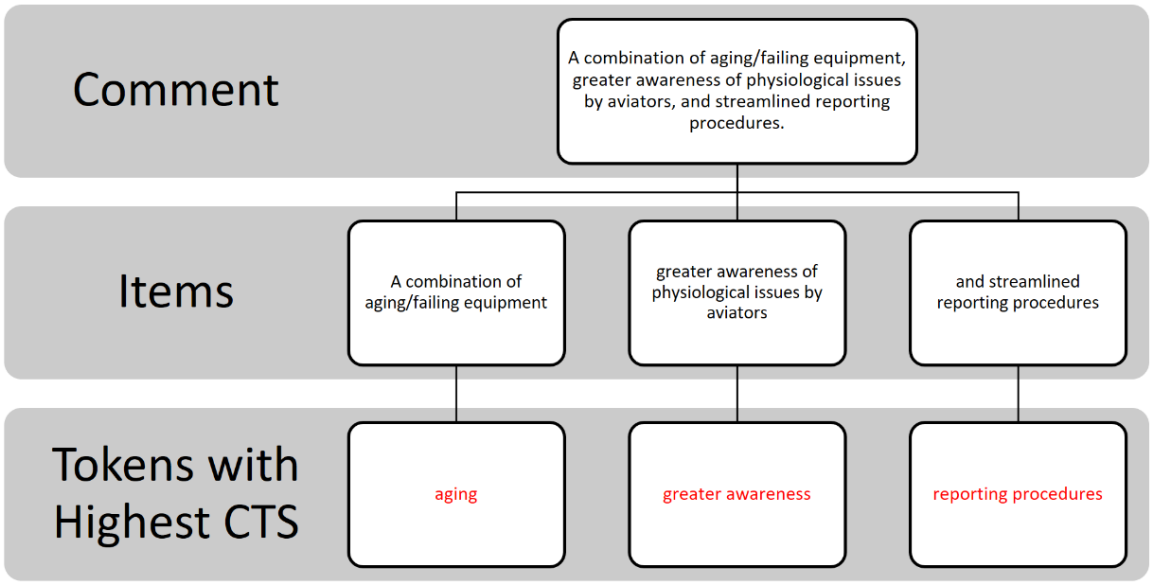

Figure 2. Breakdown of a Comment in a Non-list Format

a. Token Specific Variables

Consistent with the approach of Cairoli (2017), we describe token specific variables to be "unique to each candidate token and independent of the comment associated with the token. These variables are calculated once for each unique candidate token in the corpus and are then factored into the final CTS computation." Our approach has three token specific variables: TS, token size; TT, technical term; and PTT, partial technical term. These variables are described in detail in this section.

(1) Token Size

The variable token size (TS) is a categorical variable indicating if the token is a unigram, bigram or trigram. This captures the significance that multi-word tokens play in describing an item. Many tokens in the set begin with the same word and can contain substrings of other tokens.

(2) Technical Terms and Partial Technical Terms

The variable technical term (TT) is a binary variable indicating whether a token is a technical term or not. Technical terms such as "aerodynamic overstress," "atmospheric interference" or "environmental damage" are informative within a comment and thus offer

value when determining descriptive labels. For text analytic purposes, a technical term, according to Chuang et al. (2012b), is a multi-word phrase that meets a specific pattern. Like Cairoli (2017), we adapt the definition of a technical term to follow this pattern: "it begins with either an adjective or noun, strings together adjectives, nouns or prepositions in the middle and ends in a noun."

After tagging each word in the comment with parts of speech (POS), we use our defined TT pattern to generate a list of all TT in the comment corpus. The variable TT takes value 1 if a candidate token is in this list of TT and 0 otherwise.

A partial technical term (PTT) is a substring of a technical term. In our model, PTT is a binary variable indicating whether a token is a partial technical term or not. Unlike the previous approach where the candidate tokens used to define PTT are extracted from a combination of a reference corpus and the comment, we eliminate the reference corpus and use the strict definition of PTT. More discussion of this modification is provided in Chapter IV. The variable PTT takes value 1 if the candidate token is a substring of a TT in the TT list and 0 otherwise.

b. Comment Specific Variables

Comment specific variables are unique to each comment and are factored into the CTS computation. We use the same three comment specific variables as Cairoli (2017) and describe them here for completeness. The comment specific variables are: "Freq, the frequency of the token in each comment; RFO [relative first occurrence], a measure of the first occurrence of token relative to a token of the same frequency; FH [first half] an indication of whether a token is contained in the first half of [an item] or not" (Cairoli, 2017, p. 11).

(1) Frequency

We take the comment items corpus and our candidate tokens to construct a document term matrix (DTM). The DTM has one row per item and one column per candidate token; it stores the frequency count for that token by item. This token frequency count by item will always be greater than or equal to one. Since our comments are relatively

short, however, important tokens may appear only once. Similar to Cairoli's (2017) approach, we take the log of the candidate token frequency as the variable for computing the CTS.

(2) Positional Elements

Chuang et al. (2012b) show that the position of a token in reference to the length of the document can be useful in finding descriptive labels among the tokens. Tokens mentioned toward the beginning of a document tend to be more important, but not as important if the token appears too frequently later in the document. Absolute first occurrence (AFO) is a normalized measure between 0 and 1 of the location of a token's first appearance in a document. Similar to Chuang et al. (2012b), we calculate AFO using the normalized position of the first word in the phrase and the total number of words. As in the approach of Cairoli (2017), for bigrams and trigrams we count the *n*-grams as a single "word" in this calculation. This ensures bigrams and trigrams that begin in the same location do not have the same absolute first position. The only exception is for the tokens that begin the comment or item.

The relative first occurrence (RFO) is derived from AFO. RFO "measures 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 token in the document, then

$$RFO = (1 - AFO)^k \tag{2.1}$$

Another positional comparison we make is whether a token occurs in the first half of an item. Chuang et al. (2012b) indicate that tokens in the first sentence are more important and often better descriptors than tokens later in the document. To compensate for our short comments, we define the variable FH as a binary variable indicating whether a token appears in the first half of the item.

c. Candidate Token Score Calculation

We calculate the CTS using estimated regression coefficients similar to the approach of Chuang et al. (2012b). Using the essence of their work, we randomly select

180 comments from each of our surveys that have more than five words. We read each comment, choose a 1- to 3-gram consecutive-word token that best describes the comment and store it as an expert label. In addition, for each comment we randomly select ten tokens (excluding the expert label) to use as false-positives. This produces a data set of 3,960 comments with expert and randomly chosen labels. We also combine this data set with Cairoli's (2017) data set of 2,200 comments with expert and randomly chosen labels. We compute the comment and token specific variable values for each label-comment pair and fit a logistic regression where the response variable is 1 if the label from the comment is an expert label and 0 otherwise to estimate a new set of regression coefficients. These coefficients are shown in Table 1.

The TS variable, treated as a three level categorical variable, is represented by two binary variables whose estimated coefficients are given in Table 1. The coefficient "TS-2" in Table 1 corresponds to the binary variable that is 1 if TS is 2 (bigram) and 0 otherwise; similarly, "TS-3" corresponds to the binary variable that is 1 if TS is 3 (trigram) and 0 otherwise. The other variables are numeric, (e.g., log(Freq)) or are binary variables representing two-level categorical variables, (e.g., TT and PTT). Also included in Table 1 are standard errors and p-values for the Wald test for the corresponding coefficients. These are included as descriptive statistics as no attempt is made to formally fit a logistic regression model with appropriate model diagnostics.

Table 1. Regression Coefficients for Candidate Token Score Calculation

Model Variable	Coefficient Estimate	Standard Error
(Intercept)	-1.310 ***	0.284
TS - 2	-0.882 ***	0.260
TS - 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 = ***: $p < 0.001$, **: $p < 0.1$

5. Item Labels

The candidate tokens for each candidate item are scored using the coefficients from Table 1. The candidate token with the maximum CTS among candidate tokens for a candidate item is the label for that item. The collective unique item labels are the labels for that comment.

B. GROUP ITEMS INTO BINS

With all the item labels, we now begin the process of finding meaningful topics among the labels and sorting the items into topic bins. Before preprocessing the item labels or determining any topic bins, we modify the Cairoli (2017) approach by providing a tool that automatically constructs a list of potentially meaningful keywords to assist the sorting process. With minimal analyst review, we finalize a list of keywords (here, a "keyword" may be a unigram, bigram or trigram) that are associated with topic bins in the topic bin key for the item assignment process.

1. Create Initial Topic Bin Key

To generate an initial list of keywords for the topic bin key, we use a systematic approach that evaluates the relationships among the item labels. We create our initial list in three ways. The slight variations in the three methods maximize topic discovery to find unique keywords. The first method extracts the most frequent bigrams and trigrams from

the labels. The second method evaluates the unigrams and bigrams from item labels, but also fits an LDA and uses a saliency measure to identify phrases or words that are important, but not overly frequent. The final method uses a network approach to evaluate the correlation within a network between frequent, salient and distinct unigrams from the labels. This method also takes into account the words found in the survey question and repeated in the comment. All three methods are described in detail in this section. The resulting 1- to 3-grams from these methods become keywords that provide a starting point for determining topic bins.

a. Bigram and Trigram Frequency

The first method is the simple and straightforward approach. We tokenize the item labels into bigrams and trigrams and extract the bigrams and trigrams that do not contain stop words. The only exception for allowing stop words is if a stop word appears in the middle of a trigram. This retains keywords such as “age of aircraft,” “increase in reporting” and “lack of communication.”

The remaining bigrams and trigrams are sorted by frequency. We adjust each frequency count of these bigrams and trigrams until we reach a reasonable list of keywords. We store these keywords in the topic bin key. Figure 3 illustrates this process.

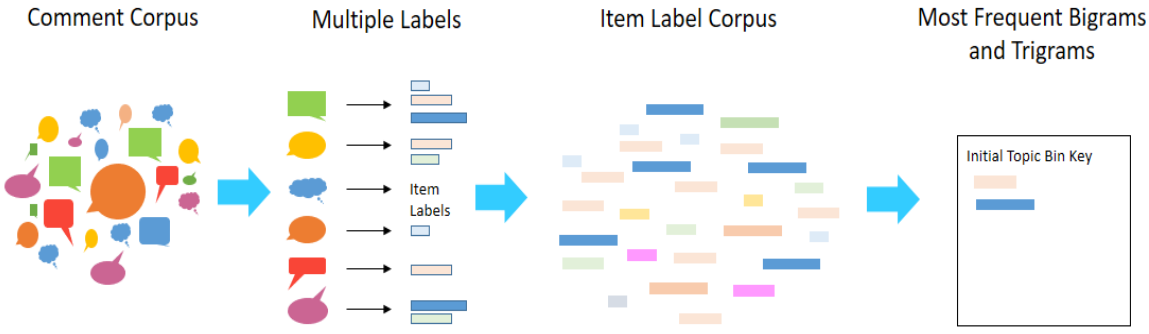

Figure 3. Frequent Keyword Extraction Process

b. LDA and Saliency

This second method of keyword extraction is more involved but finds keywords the first method does not. We create a label corpus and form two separate DTMs: one with the item labels tokenized into bigrams and the other as a standard DTM with unigrams. As in Cairoli's (2017) approach, the common topic modeling technique LDA is useful on our item labels to determine topics, or in this case, keywords. We fit an LDA model to determine a number of topics. LDA identifies the latent topics *T* of a corpus of documents and allows each document *d* to correspond to multiple topics. From the LDA fit, we can estimate the distribution of words (or tokens) *w* in each topic, $\{p(w|T)\}$ as well as the distribution of topics for each document, $\{p(T|d)\}$ (Silge & Robinson, 2017). These distributions allow us to extract the distinct and salient words in the item labels.

As defined by Chuang et al. (2012a), 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, $\{p(T)\}$. Using Bayes rule to estimate $\{p(T|w)\}$ and the label DTM to estimate $\{p(T)\}$, *distinctiveness* is defined as:

$$distinctiveness(w) = \sum_T p(T \mid w) \log\left(\frac{p(T \mid w)}{p(T)}\right) \tag{2.2}$$

Chuang et al. (2012a) use *saliency* as a measure to find relevant, but not overly frequent words in topics. Saliency is the product of the probability of selecting a word or token *w* from the corpus of words, *p(w)* and the distinctiveness. The saliency of a word is defined as:

$$saliency(w) = p(w) * distinctiveness(w) \tag{2.3}$$

Fitting the LDA model on our label DTMs, we are able to extract the most salient bigrams and unigrams from the labels. We add any new bigrams and unigrams discovered from this method to our list of keywords in the topic bin key. Figure 4 illustrates the process of extracting salient keywords.

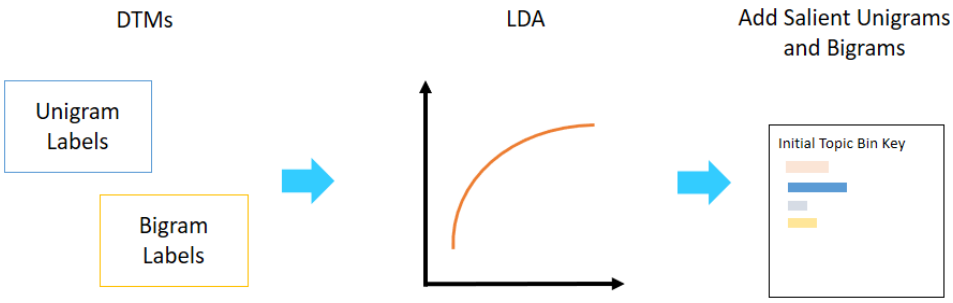

Figure 4. Salient Keyword Extraction Process

c. Networks

The third method of extracting keywords involves constructing a network of the unigrams within the item labels. With the frequent, salient and distinct unigrams determined from the labels and LDA fit, we evaluate the correlation between all said terms. We also consider the words found in the survey question. The responses in surveys sometimes begin with repeating the first part of the question. Without considering the words in the original survey question, many of those terms become either labels or keywords and do not usually provide us with meaningful information. As a result, we take into account these question words by increasing the correlation threshold for those words. The correlation for combinations that include question words must be higher to be considered a "good" keyword or topic.

From the terms of interest, we construct a non-directed network where the nodes are the unigrams and the arc weights represent the correlation between words. For a given correlation threshold (where arcs with correlation below the threshold are broken), two-way paths and three-way cycles within the network become keywords. Figure 5 illustrates this process. In this network method, there are redundancies where phrases similar in concept remain and lexical variations, where the order of the words are reversed, are likewise retained. For example, in Figure 5 the keywords "3-4" as well as "4-3" are added to the topic bin key even though they are mere lexical variations of each other and are likely conceptually similar. Other examples of these redundancies occur for stemmed words or synonyms. For the purposes of the visualization, redundancies are removed but term grouping or redundancy reduction is required here for constructing a more accurate and concise initial topic bin key.

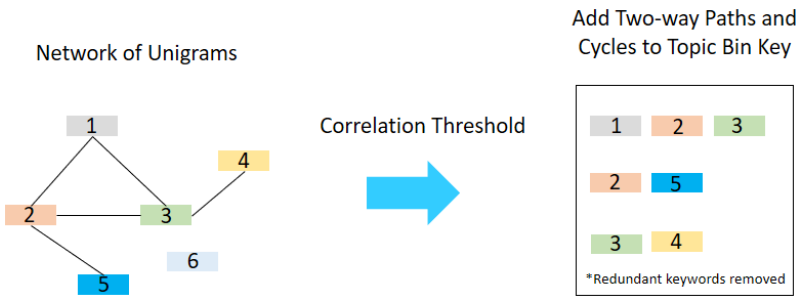

Figure 5. Correlated Keyword Extraction Process

2. Finalize Topic Bin Key

With an initial list of keywords, we (the analysts) continue to review visual plots such as correlation plots or word clouds, to provide any additional keywords based on the comment question. Here, we use any background knowledge of the survey subject along with subject matter expertise to verify that the topic bins make sense and are meaningful.

3. Assign Items to Topic Bins

The slight modification in our method is that we bin items instead of each comment. Since comments in our methodology can have multiple items and consequently labels, there are more topics per comment to bin and quantify. The algorithm of assigning items to topic bins using the topic bin key is similar to the comment binning of Cairoli (2017).

> The topic bin key is first compared to the [item] labels. Using regular expressions to allow for partial matches, labels are searched for each keyword from the topic bin key and the positions of the matches are saved and compared. The match that appears earliest in the label is considered the primary topic and the label is assigned to that keyword's corresponding topic bin. For labels that do not contain any keyword matches, the comments are reviewed to determine if they contain keywords. Matches are assigned to corresponding topic bins. (Cairoli, 2017, p. 27)

A topic bin frequency table is reviewed to determine if binning any remaining items as "Other" would be acceptable. This binning assignment process continues until an appropriate number of items are assigned to topics.

III. SURVEY RESULTS

The following sections detail the backgrounds of the two surveys we analyze in this research. While both surveys are Department of the Navy (DoN) related, they differ both in subject matter and the style of language used in the text responses. Applying our methodology to responses from two very different surveys is helpful for assessing how robust our comment analysis approach is and for identifying potential difficulties or areas for modifications.

A. DEOCS

The first survey is an organizational assessment called the Defense Equal Opportunity Management Institute (DEOMI) Organizational Climate Survey (DEOCS). This survey was administered to a Navy command to garner members' perceptions of their command's climate. Text comments from these types of surveys tend to be rather informal and topics vary significantly. In this section, we provide more background on this survey and describe how we implement our methodology on an example response with illustrations of each step.

1. Background

This research extends the process of identifying prevalent topics for text responses to the questions of interest to the Naval Inspector General (NAVINSGEN). The NAVINSGEN mission is "to inspect, investigate or inquire into any and all matters of importance to the Department of the Navy" and to strive to maintain the highest level of public confidence (Department of the Navy [DoN], 2005). NAVINSGEN "conducts inspections and surveys, making appropriate evaluations and assessments concerning operating forces afloat and ashore, DoN components and functions and Navy programs, which impact readiness or quality of life for military and civilian naval personnel" (DoN, 2005).

One method NAVINSGEN uses for organizational assessment is the DEOCS. These command climate assessments must be completed within 90 days of a Commanding

Officer (CO) taking command and annually thereafter "to determine the "health" and functional effectiveness of an organization by examining such factors as morale, teamwork and communication" (DoN, 2017).

Individual commands and NAVINSGEN review the surveys, but often overlook or require extended periods of time to analyze the comments, which may reduce their relevance and the command's ability to make timely changes. Since comments are more informative when and where there is a command climate issue, automating a process to identify prevalent topics and offer sample comments can provide invaluable insight to commands and their leadership to find issues before they lead to significant problems.

The DEOCS data we use comes from a Navy command and contains nine sections that contain free-text comment boxes. We use a sample from the combined 960 responses in this survey to contribute to the comment-label pair data set used for estimating the regression coefficients. The analysis in this section, though, will only focus on the 149 text responses to one question from this survey: "What is one thing your command can do to reduce your stress level?"

2. Comment Analysis Application

A sample comment will be used from the DEOCS to demonstrate how we obtain a descriptive label for each comment.

Example Comment:
Clear and concise direction follow written procedure.

a. Preprocess Candidate Items

Because this particular comment is not in a list format, nor does it contain any commas, it becomes the sole candidate item and is not parsed further into several candidate items. Items are imported into the statistical computing environment R (R Core Team, 2017) and are converted to lower case, replacing contractions with non-contraction equivalents and removing punctuation.

Preprocessed Item:
clear and concise direction follow written procedure

b. Candidate Tokens

We use the function `DocumentTermMatrix()` from the R package **tm** (Feinerer & Hornik, 2017) to construct all unigrams, bigrams and trigrams from the candidate item and store their document frequency count in a DTM. Table 2 shows the set of candidate tokens for the example comment, with any tokens of a single stop word removed. The variable calculations in the next step help us determine which of these tokens can best describe the example comment.

Table 2. Candidate Tokens

and concise	direction follow
and concise direction	direction follow written
clear	follow
clear and	follow written
clear and concise	follow written procedure
concise	procedure
concise direction	written
concise direction follow	written procedure
direction	

c. Variable Calculations

This section explains the process for each variable value calculation.

(1) Token Size

Using regular expressions, we determine the number of words contained in each token. Table 3 displays the TS for each token.

Table 3. Token Size

and concise	2	direction follow	2
and concise direction	3	direction follow written	3
clear	1	follow	1
clear and	2	follow written	2
clear and concise	3	follow written procedure	3
concise	1	procedure	1
concise direction	2	written	1
concise direction follow	3	written procedure	2
direction	1		

(2) Technical and Partial Technical Terms

We use the R package **udpipe** (Wijffels, 2018) to identify POS elements for the corpus of comments. With our definition of a technical term, we can then generate a list of technical terms found in the comment corpus. A candidate token's variable TT is assigned value 1 if it is found in this list of technical terms and 0 otherwise. Likewise, PTT is assigned value 1 if it is a substring of a technical term and 0 otherwise.

(3) Frequency

We extract the frequency of candidate tokens for each document from the comment corpus DTM.

(4) First Half

We use the `gregexpr()` function to determine the total number of words in the item. We divide this number by 2, rounding up to the nearest whole number, to determine the cutoff between the first and second half of the item. Using the `strsplit()` function, each item is truncated at the cutoff position and the first half-item is stored. We compare the candidate token to the half-item to determine if it appears in the first half. The variable FH takes value 1 if the entire token is contained in the first half and 0 otherwise.

Example First Half:
clear and concise direction

d. Candidate Token Score Calculation and Labels

We calculate the CTS using the sum product of the regression coefficients in Table 1 and the item's variable values. The variable values for the example comment are summarized in Table 4. The candidate token with the maximum CTS, "concise direction," in this example, is assigned to be the label for the comment.

Table 4. Variable Summary with Final Candidate Token Score

Coefficient	**1**	**2.313**	**-0.023**	**0.918**	**2.025**	**0.436**	
Token	TS	TT	PTT	log(Freq)	RFO	FH	CTS
and concise	-0.88	0	0	0	0.80	1	-0.14
and concise direction	-1.01	0	0	0	0.75	1	-0.36
clear	0.00	0	1	0	1	1	1.00
clear and	-0.88	0	0	0	1	1	0.27
clear and concise	-1.01	0	0	0	1	1	0.14
concise	0.00	0	1	0	0.67	1	0.32
concise direction	-0.88	1	1	0	0.60	1	1.62
concise direction follow	-1.01	0	0	0	0.50	0	-1.30
direction	0.00	0	1	0	0.50	1	-0.01
direction follow	-0.88	0	0	0	0.40	0	-1.38
direction follow written	-1.01	0	0	0	0.25	0	-1.81
follow	0.00	0	0	0	0.33	0	-0.63
follow written	-0.88	0	0	0	0.20	0	-1.79
follow written procedure	-1.01	0	0	0	0	0	-2.31
procedure	0.00	0	0	0	0	0	-1.31
written	0.00	0	0	0	0.17	0	-0.97
written procedure	-0.88	0	0	0	0	0	-2.19

e. Create Topic Bin Key

For the initial list of keywords, we first tokenize the collection of item labels into bigrams and trigrams using the `unnest_tokens()` function in the R package **tidytext** (Silge et al., 2016). The most frequent bigrams and trigrams are stored in the topic bin key. Figure 6 depicts the most frequent bigrams (with no stop words) from the DEOCS labels where arrows connect the two words in the bigram and indicate their order in the bigram. Filtering out *n*-grams with stop words, adjusting by frequency and manually removing redundancies produces the initial list of keywords in Table 5.

Further, we construct a corpus with the item labels using the **tm** function `VCorpus()` (Feinerer et al., 2017). We train an LDA model with the `LDA()` function from the **topicmodels** package (Grun & Hornik, 2011) and determine the "best" number of topics by locating the "knee" in a log-likelihood, plotted as a function of the number of topics. The LDA model is fit to the number of topics found and we estimate saliency for each unigram or bigram using Equations (2.2) and (2.3) in Chapter II. Any new and valuable keywords found with this method are added to the topic bin key.

For the network method, the frequent and salient terms and their correlations are used to construct a network using the packages **network** (Butts, 2015) and **igraph** (Csardi & Nepusz, 2006.) This network is used to find additional keywords as illustrated in Figure 7. For a given correlation threshold, we add bigrams and trigrams identified from two-way paths and cycles in the network to the topic bin key if they do not already appear in the key.

Figure 6. Most Frequent Bigrams from DEOCS labels

Table 5. Compiled Initial List of Keywords for DEOCS

18 months	distinct timelines	real job
5 minutes	duty in line	real plans
actual instruction	educate leaders	reliable schedule
administration paperwork	equal recognition	repair officer
aggressive threats	equipment last minute	respect of time
alternate means	eval writing	sailors musters
boot camp	family day	schedule from tycom
chain of command	gapped billets	screen jobs
cmcs comment	helpful with paper	senior chief
collateral duties	incentives for people	sports day
command functions	involved activities	ssdf mix
command picnic	lower chain	st po
command pt	lpo billet	stress management
command sports	morale boosters	stressful job
comrel opportunitites	mre boxes	submarine personnel
concise direction	navy minimum	support services
concrete plan	passive aggressive threats	surprise time
consistent in port	positive stressors	tad with units
date technology	providers inside	time card
definite knowledge	qualified individuals	traffic home
delinquent study	rate training	workout period

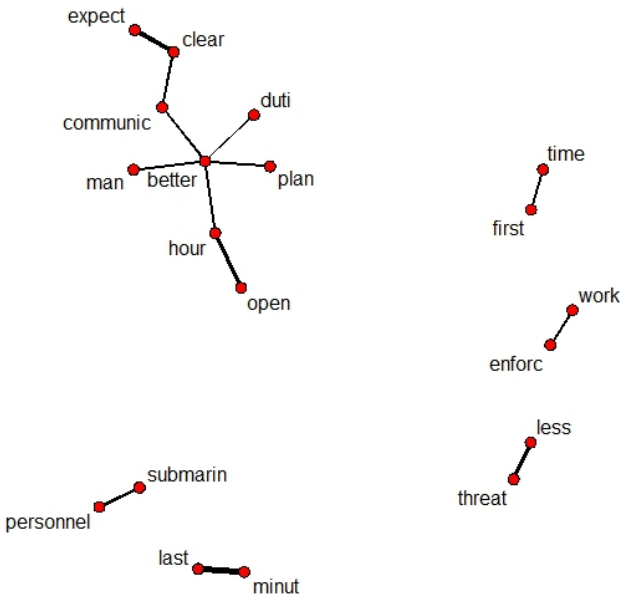

Figure 7. Network of Correlated Stemmed Unigrams from DEOCS Labels

f. Assign Items to Bins

Using our subject matter expertise review, we determine ten prevalent topics to bin the 288 items from the 149 DEOCS comments for the question "What is one thing your command can do to reduce your stress level?" The ten topics are *Destructive Behaviors*, *Instructions*, *Job*, *Leadership*, *Morale Welfare and Recreation* (MWR), *Recognition*, *Schedule*, *Sports Day*, *Stress Management* and *Technology*. Table 6 summarizes the topic bin key and lists the prevalent topics for this survey question with example keywords.

Table 6. Summary of DEOCS Topic Bin Key

Topic	**Example Keywords**
Destructive Behaviors	hostile work, discriminatory, aggressive threats
Instructions	knowledge, study,
Job	eval writing, collateral duties, billet, duty in line, rate training, qualified individuals,
Leadership	chain of command, command master chief , commanding officer, responsibility, decision
MWR	command picnic, family day, morale boosters, liberty
Recognition	equal recognition, reward, incentives
Schedule	long hours, plan, last minute, time card, distinct timelines
Sports Day	command sports day, pt, workout
Stress Management	support services
Technology	tablets, tools, Internet

Figure 8 displays the proportion of items that pertaining to each of the determined prevalent topics for the question "What is one thing your command can do to reduce your stress levels?" The top three topics are *Job*, *Schedule* and *Leadership*. As implied by the keywords in Table 6, the topic *Job* pertains to anything fundamentally related to a sailor's work requirements. Likewise, *Schedule* pertains to themes related to time while *Leadership* pertains to matters regarding the chain of command.

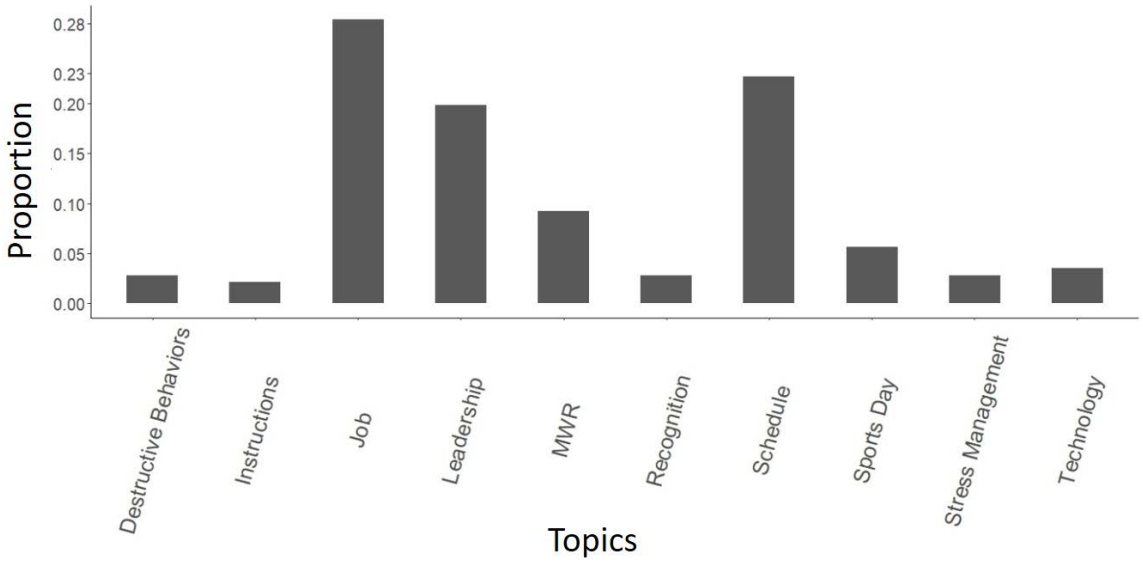

Figure 8. Binned Items from DEOCS Question

B. PHYSIOLOGICAL EPISODES SURVEY

The second survey is part of a Navy investigation to solicit information regarding the recent issues of PE within Naval Aviation. The comments in this survey are very technical and professional in nature. In this section, we provide more background and describe how we apply the comment analysis on a sample comment with examples of each step of the process.

1. Background

The second set of data reviewed in this thesis is concerned with the cause of recent PE in the Naval Aviation community. The Naval Aviation Safety Management System states a PE occurs whenever any of the following conditions exist outside of a Naval Aviation mishap:

> (1) Hypoxia, proven or suspected. (2) Carbon monoxide poisoning or other toxic exposure. (3) Decompression sickness because of evolved gas (bends, chokes, neurocirculatory collapse) or severe reaction to trapped gas resulting in incapacitation. (4) Hyperventilation. (5) SD or distraction resulting in unusual attitude. (6) Loss of consciousness for any cause. (7) An unintentional rapid decompression exposing personnel to cabin altitudes above flight level 250, regardless of whether dysbarism or hypoxia occurs.

(8) Other psychological, pathological or physical problems that manifest during or after actual flight. (DoN, 2014, pp. 109–110)

The Navy began observing an inexplicable increase in reported PE events beginning in 2009 (Physiological Episodes [PE], 2018). At one point, the concerns about PE became so prevalent and serious that aviators refused to fly until their leadership addressed the issue, which eventually led to a Naval Aviation operational pause. The Navy has since established teams such as the Physiological Episode Team (PET), and the Physiological Episode Action Team (PEAT). On February 6, 2018, Rear Admiral Sara A. Joyner testified before the House Armed Services Committee on PEs within Naval Aviation. In her statement, Rear Admiral Joyner says addressing PE remains "the number one safety priority for the entire Naval Aviation community" (PE, 2018).

In 2016, the Naval Postgraduate School conducted the Physiological Episode Survey as part of the U.S. Navy's investigation into the recent increase in PE experienced by F/A-18 and T-45 aircrews. The purpose of the survey was to solicit information from aircrews and maintainers, regardless of whether or not they had experienced a PE. The text analysis of the responses from the survey can help focus the investigation into areas that show the greatest promise for determining the root cause(s) of PE.

The PE survey contains 138 questions. One of these questions contains a text comment box for respondents to state why they believe there was an increase in PE. The survey was administered to three groups: T-45, F/A-18 aircrews and F/A-18 maintainers, resulting in 1,060 responses. Our focus is on 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?"

2. Comment Analysis Application for Multiple Labels

A sample comment will be used from the PE survey to demonstrate how we obtain multiple labels from a listed comment.

Example Comment:

1. More sensitivity to it in our community. 2. More requirements to report it. 3. Young aviators (potentially) not knowing the difference between not being physically prepared/ adapted vs hypoxic

a. Preprocess Candidate Items

Before preprocessing, we first parse the comment into candidate items corresponding to the list. Preprocessing involves converting to lower case, removing punctuation and replacing contractions with the non-contraction equivalent. This particular comment is parsed into three items. Table 7 displays how this comment is itemized and preprocessed.

Table 7. Preprocessed Comment Items

more sensitivity to it in our community
more requirement to report it
young aviators potentially not knowing the difference between not being physically prepared adapted vs hypoxic

b. Calculate Variable Values and Compute CTS

For each candidate item, we tokenize and calculate variable values using the steps in Chapter III, Section A, Subsections 2a-d. The candidate token in each item with the maximum CTS become the labels for this comment example. For this comment, the labels are “more sensitivity,” “more requirements” and “young aviators.”

c. Create Topic Bin Key

For the initial list of keywords, we first tokenize the collective item labels into bigrams and trigrams. The most frequent bigrams and trigrams are stored in the topic bin key. Figure 9 depicts the frequent bigrams from the PE survey item labels. We construct a corpus with the item labels to train an LDA model and determine the “best” number of topics by locating the “knee” in a log-likelihood plot. The LDA model is fit to the number of topics found and we estimate saliency for each unigram and bigram. Any new and valuable keywords found with this method are added to the topic bin key. Filtering out stop words, adjusting by frequency and removing redundancies produces the initial list of keywords in Table 8.

For the network method, the frequent and salient terms and correlations are used to construct a network to find additional keywords as illustrated in Figure 10. For a given correlation threshold, we use the two-way paths and three-way cycles in the network to find bigrams and trigrams to add to the topic bin key if they do not already appear in the key. We review the initial list of keywords to manually remove redundancies or lexical variations. We also add missing keywords using subject matter expertise to finalize a topic bin key.

Table 8. Initial List of Keywords for PE Survey

aging equipment	ecs components	obogs systems
aging systems	flight hours	physiological episodes
aircrew awareness	hyper sensitivity	robd training
cabin pressure	increased awareness	
cabin pressurization	legacy hornets	

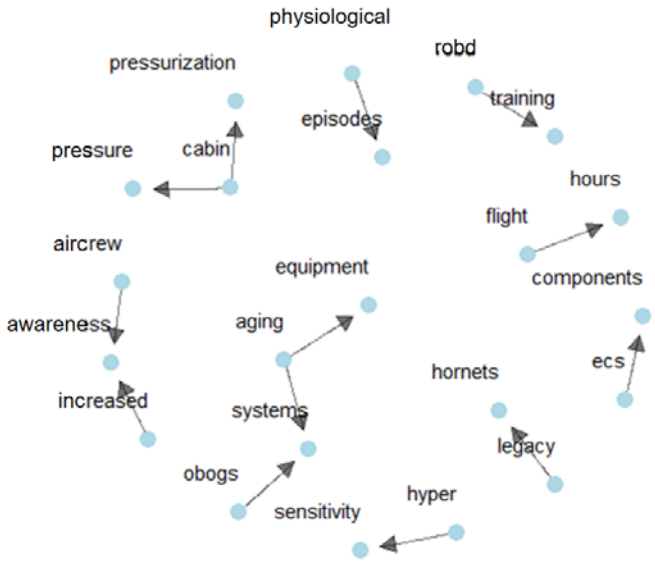

Figure 9. Most Frequent and Salient Bigrams from PE Labels

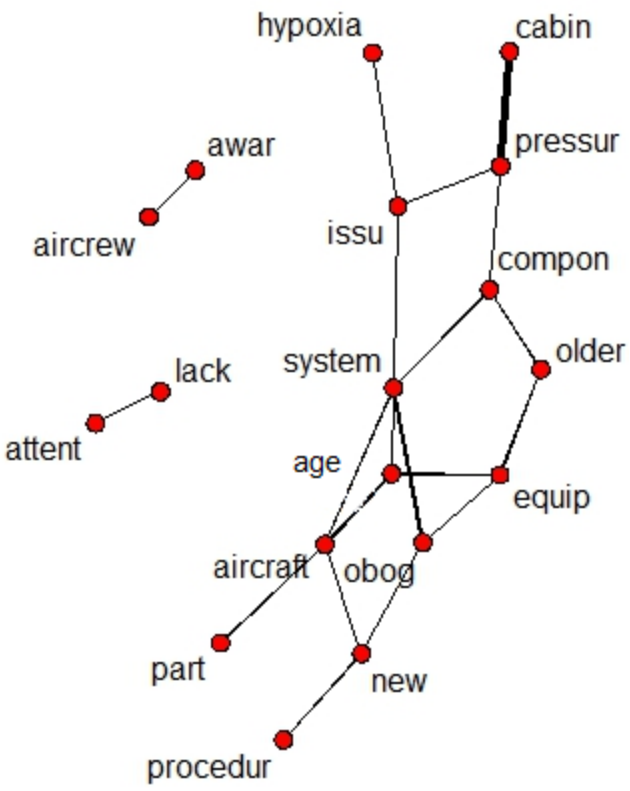

Figure 10. Network of PE Stemmed Labels

d. Assign Items to Bins

Again, appealing to our subject matter expertise review, we construct nine prevalent topics to be used to bin the PE items. The nine topics are *Age*, *Aircrew*, *Awareness*, *Environment*, *Funding*, *Leadership*, *Maintenance*, *Parts* and *PE Criteria*. Table 9 summarizes these topics and some pertinent corresponding keywords. The comments from this survey yield 2,872 items. Figure 11 displays the proportion of these items in each of the prevalent topics.

Table 9. Summary of PE Survey Topic Bin Key

Topic	Example Keywords
Age	old, legacy, hours
Aircrew	training, experience, lack of knowledge, diet, sleep
Awareness	heightened, hypersensitivity, recognition
Environment	temp flow, air quality
Funding	money, budget
Leadership	lack of action, mismanagement, navy
Maintenance	care, fix, maintainers
Parts	equipment, obogs, design, material, system
PE Criteria	symptoms, definition, criteria

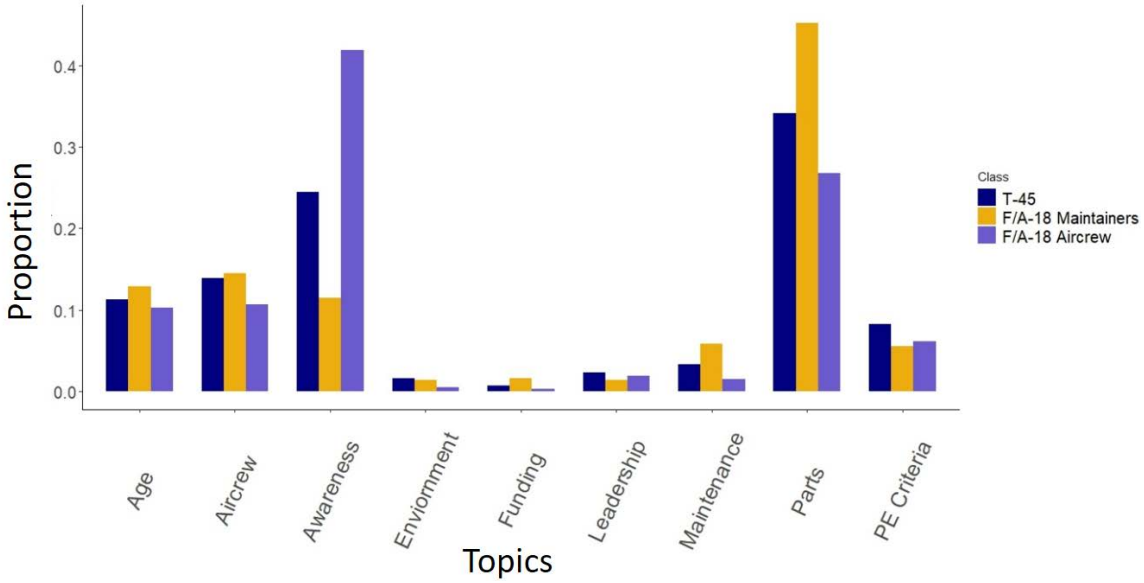

Figure 11. Binned Items from PE Survey

From Figure 11 and Table 9, we see that the most frequent response to the question regarding the reasons for an increase in PE pertain to specific aircraft parts or equipment and to an increase in awareness. Reasons pertaining to age, aircrew or PE criteria are mentioned less often and the other four topics are mentioned even less often. We also see differences between F/A-18 maintainers and F/A-18 aircrews. Among F/A-18 maintainers, the most frequently mentioned reason for an increase in PE pertains to specific parts or

systems of the aircraft. For F/A-18 aircrews, the most frequently mentioned reason for the PE is due to an increase in awareness of the issue. The ability to classify items according to topic along with a topic bin key allows the analyst to quantify text responses. As in the PE survey, this is particularly important when polling diverse populations to uncover differences of opinion or unique viewpoints.

IV. DISCUSSION

In this Chapter, we provide more in-depth explanations of the modifications to the comment analysis approach and the motivations behind them.

A. GENERALIZED MODEL

The method of estimating regression coefficients for the CTS calculation uses the essence of the work of Chuang et al. (2012b) and Cairoli (2017). It involves fitting a logistic regression to classify a comment's candidate label as a good or "expert" label for that comment or not. The estimated coefficients are then applied to each candidate token's variable values to compute a candidate token score. The estimated coefficients used for the DEOCS and PE survey and their standard errors are reproduced in Table 10 under *Generalized Model*. Table 10 also displays the estimated coefficients used in Cairoli (2017) which are derived from comments in surveys regarding Navy retention and uniforms. Cairoli's coefficients include the Reference Commonness (RC) categorical variable, which we omit in our generalized model (discussed in Chapter IV Section B).

Table 10. Estimated Regression Coefficients.
Adapted from Cairoli (2017).

	Generalized Model		***Cairoli (2017)***	
Model Variable	Coefficient Estimate	Standard Error	Coefficient Estimate	Standard Error
(Intercept)	-1.310 ***	0.284	-2.522 ***	0.625
TS - 2	-0.882 ***	0.260	-1.125 **	0.409
TS - 3	-1.005 ***	0.278	-1.281 **	0.451
TT	2.313 ***	0.321	3.293 ***	0.568
PTT	-0.023	0.260	-1.048*	0.438
Log(Freq)	0.918 •	0.501	0.574	0.827
RFO	2.025 ***	0.405	3.801 ***	0.827
FH	0.435 •	0.262	0.330	0.508
RC ϵ (0%, 20%]	-	-	-0.278	0.563

	Generalized Model		***Cairoli (2017)***	
Model Variable	Coefficient Estimate	Standard Error	Coefficient Estimate	Standard Error
RC ϵ (20%,40%]	-	-	-0.829 •	0.457
RC ϵ (40%, 60%]	-	-	-0.531	0.446
RC ϵ (60%, 80%]	-	-	-0.185	0.503
RC ϵ (80%, 100%]	-	-	-2.924 *	1.170

Statistical significance = ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, •: $p < 0.1$

Like Cairoli (2017), we estimate a set of coefficients to apply to comments for two surveys. We randomly select 180 comments from each of our two surveys (approximately 10% of the total number of text responses in our surveys) that have more than five words. For each of those comments, we apply a 1- to 3-gram consecutive-word label and store it as an expert label. In the logistic regression, the response variable is 1 if the token is an expert label and 0 for 10 randomly selected tokens (excluding the expert label), that we generate as false-positives and give a weight of 0.1. This produces a data set of 3,960 comments with labels.

This method of estimating the regression coefficients works well in producing descriptive labels both in our survey data and in that of Cairoli (2017). However, reading even a subset of comments and hand labeling each comment can delay analysis. The natural response to streamline this portion is to determine how few comments are required to estimate valuable coefficients, but this "minimum number" is dependent on the question and survey. Our experiments show that only about 10% and 25% of the comments need to be read and assigned an expert label for the PE and DEOCS surveys, respectively. Rather than find the minimum number of comments necessary to produce effective estimated coefficients for a given survey, we combine our hand-labeled comments with the set of 2,200 label-comment pairs from Cairoli (2017). The intent of combining the data from four surveys is to produce coefficients that are more stable. The topics among all the combined comments are broad enough that these coefficients should be appropriate for any set of survey comments, regardless of survey topic.

With the combined comment-label pairs (560 total) and subsequent logistic regression coefficients given in Table 10, we assess how well the logistic regression is predicting the correct token as the expert label. We assess the accuracy of our logistic regression fit by determining for each token an estimated probability that the token is the expert label. For each comment, there are 11 tokens, the actual expert label token and 10 false-positive tokens. Hence, there are 11 estimated probabilities per comment. We rank the 11 estimated probabilities and store the rank pertaining to the actual expert label. For all the comments used in our logistic regression data, 45% of the actual expert label tokens rank first among the 11 within its respective comment's tokens. 80% of the expert label tokens rank in the top four positions. Table 11 gives a breakdown of the proportion of expert label tokens that fall in the top one through six ranks.

Table 11. Accuracy of Logistic Regression

Ranks	Proportion
Top 1	0.452
Top 2	0.601
Top 3	0.721
Top 4	0.801
Top 5	0.865
Top 6	0.897

B. REFERENCE CORPUS

The approach described in Chapter I, Section B relies on a valid external reference corpus for establishing a basis of the language or prevalence of technical terminology in a survey. The reference corpus "can be the document that the token comes from, the entire corpus of documents being labeled or an entirely separate corpus created from general web scraping" (Cairoli, 2017, p. 10). The original approach uses a reference corpus for two reasons. The first is to provide a measure of commonness, which can be a good indicator of a descriptive label. This measure of commonness, known as Reference Commonness (RC), is a token specific variable used in the CTS calculation. RC is "calculated for each 1- to 3-gram token contained in the reference corpus by dividing the log of the token

frequency in the reference corpus by the log of token frequency of the most frequent token of the same token size, where RC=0 for a token that does not appear in the reference corpus" (Cairoli, 2017, p. 11). The second reason for a reference corpus is to provide a single source for identifying partial technical terms. This is to ensure a consistent definition of PTT across time, survey or subset of survey responses.

Although it is straightforward to identify a good reference corpus for the PE survey, it is more difficult to identify one for the DEOCS that corresponds to its more informal language and broad range of topics. Because there is no clear choice of reference corpus for this survey, we use the pooled comments from all nine questions with free text responses in the DEOCS as the reference corpus. This approach produces reasonable labels. However, the process of preparing the reference corpus can be laborious.

The reference corpus preparation involves multiple steps. First, the process involves removing the corpus of redundant or extraneous text like headers or footers. In addition to the standard preprocessing such as converting to lower case and replacing contractions, we tokenize the reference corpus into unigrams, bigrams and trigrams. For each token, we calculate its RC. Further, we generate a list of all technical terms form the reference corpus. Implementing these steps and using the reference corpus of pooled comments (when a reference corpus cannot be identified) produces reasonable labels, but we demonstrate next eliminating the process altogether still produces descriptive labels and simplifies the entire comment analysis process.

To demonstrate the effect of omitting the reference corpus, we replicate one of our experiments here. In this experiment, we focus on the PE survey and generate two models: one with a reference corpus and one without. To estimate coefficients for these models, we use a data set of 180 randomly selected and hand-labeled comments. To compare these two models, we apply each of the models to separate data set consisting of 102 comments, randomly selected from the PE comments and excluding the 180 used to estimate coefficients. We look at three metrics based on the labels generated, the differences in computation time and the number of PTT generated.

Model 1 requires a reference corpus, which we select to be the Naval Aviation Safety Management System instruction (DoN, 2014), specifically Chapters 3 and 5, which detail Mishap and Injury Classification and Hazard Reports. With this reference corpus, we remove the recurring header and date which appear on every page of the document, since this repetition could affect our RC calculation. The reference corpus is stemmed and tokenized and we calculate the RC variable for every 1- to 3-gram token, by taking the log of the frequency of that token divided by the log of the most frequent token of the same size. From these reference corpus tokens, we also extract the tokens that follow our defined technical term pattern. This list of technical terms derived from the reference is added to the list of technical terms generated from each comment for a total of 16,609 technical terms. For Model 1, we use this pooled list to determine whether a token is a PTT or not. Model 2 does not require a reference corpus, thus we do not calculate the RC variable, and we extract technical terms only from the comment context.

Of the 102 labels generated from each model, 75 are exact matches. The 27 non-matches are displayed in Table 12, with bold words indicating the tokens that appear in both labels. Among these 27 non-matching pairs, 15 are at least a substring of the corresponding label from the other model. Thus, in this experiment, by removing the reference corpus there are only 12 labels out of the 102 comments that are completely different. Further review of these non-matching labels shows that the labels generated from Model 2, the model *without* a reference corpus, tend to be more descriptive than labels from Model 1.

Table 12. Non-matching Labels from Model 1 and Model 2

Model 1 Labels	**Model 2 Labels**
jet issues	combination of **jet**
airframes	**airframes** are
members being	them a bit
awareness	**awareness** and
beyond	increased awareness
sensitivity	increased **sensitivity**

Model 1 Labels	Model 2 Labels
aircraft	**aircraft** and
think	occasional issue
aircrew training	awareness of pes
sieve beds	charcoal **sieve beds**
think	aircraft cause
felt	aircrew finally
irt	canopy seals
always	they have
aging	**aging** equipment
think	that we
maybe	chemical change
significant pes	couple **significant pes**
increased	**increased** knowledge
legitimate pe	few **legitimate pe**
getting	obogs system
aircraft systems	age of **aircraft**
we have	mishap that
minor episode	more **minor episode**
low sa	student **low sa**
awareness	**awareness** has
i do not	**i do**

Model 1 uses a reference corpus, Model 2 does not

Applying the two models to this data set of 102 comments, we also measure the time spent calculating the token specific variables, since this is the only step in the algorithm that the reference corpus would affect. While the computation time will depend greatly on the efficiency of the machine, as a baseline, the time for Model 1 to calculate the token specific variables is 278 seconds and the for Model 2 is 21 seconds. Without the reference corpus, the computation time is an order of magnitude shorter and this amount of time saved does not include the time it takes to identify, acquire, produce and preprocess the reference corpus.

As far as PTT comparison between models, Model 1 generated 1,118 unique PTTs while Model 2 generated 963 unique PTTs. In this experiment and since Model 2's PTT is a subset of Model 1's, the reference corpus only provided an addition 155 PTTs.

The results of this experiment and many more similar to it, leads us to conclude that omission of the reference corpus may not be needed for surveys with a sufficient number of comments. There may be cases, though, that a reference could still be helpful, particularly for smaller data sets of less than 1,000 comments. In cases like these, articles from the newspaper the Navy Times could potentially be useful as a reference corpus since there are articles that correspond to virtually any DoD survey topic.

C. THE EFFECT OF ITEMIZING AND MULTIPLE LABELS

To demonstrate the effect of itemizing the comments before applying the algorithm, we apply our method with and without itemizing to the same 102 comments used in Chapter IV Section B. In this experiment, with itemizing we expect to have a greater number of labels overall, but we also hope to capture quality labels that otherwise would have been missed without the initial parsing step. For this set of 102 comments, the method using multiple labels generated an extra 156 labels. We remove any stop words or words such as "however," "also" and "although" from this list end up with 101 new yet quality labels generated simply by initially parsing the comment into items.

The multiple-labels method better matches human responses and captures more ideas or topics from the entire comment since it is not limited to a single 1- to 3-gram token. For example, the example comment from Chapter III yields these labels using the multi-label approach.

Example Comment:

1. More sensitivity to it in our community. 2. More requirements to report it. 3. Young aviators (potentially) not knowing the difference between not being physically prepared/adapted vs hypoxic

The three labels "more sensitivity," "more requirements" and "young aviators" describe the entire comment adequately and succinctly for easy and more accurate analysis.

Another added benefit to initially parsing by items is that this method provides more accuracy when quantifying prevalent topics. In this experiment, for the single-label method, labels pertaining to the topic of "age" occur 9 times. In the multiple-label method, the topic of age occurs 13 times. Table 13 shows one case in which 'aging' occurs as a label when itemizing, but not in the single-label algorithm. Table 13 also gives other example comments comparing the labels we obtain from the multiple-label method to the single-label method. In this experiment, the multi-label approach labels are more descriptive and include topics that might otherwise have been overlooked.

Table 13. The Effect of Itemizing

Comment	Single Label (no itemizing)	Multiple Labels (with itemizing)
Age of aircraft, lack of PMs on system, rare chemical reactions that occurs with some other, yet identified, factor(s).	• age of aircraft	• age of aircraft • lack of pms • rare chemical reactions
Some due to systems aging, however most are due to people attributing any possible issue, such as disorientation to OBOGS, where it more likely is often a cold, diet, sleep, unfamiliarity with the aviation environment, etc.	• due	• aging • such as disorientation • diet • sleep • unfamiliarity
Equipment degradations, airframe extensions, little-to-no accountability from NAVAIR to monitor the health of all the components in the jet system.	• equipment degradations	• equipment degradations • airframe extensions • accountability from
Older equipment, better awareness.	• older equipment	• older equipment • better awareness
Age of aircraft systems, primarily ECS and OBOGS. Misunderstanding of the periodicity of cleaning and upkeep required to keep the systems healthy in the long term, such as parts replacement.	• age of aircraft	• age of aircraft • primarily ecs • periodicity of cleaning • such as parts

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

The methodology described in this thesis builds on the foundation of comment analysis in DoD surveys developed by Cairoli (2017). For open-ended survey questions, where the responses provide prospects for rich qualitative data, researchers now have an enhanced method of quantifying the results and viewing prevalent topics. Analysis of open-ended survey comments afford more opportunities to gain insights into unfamiliar topics or discover information not found in closed-ended questions. This not only allows for findings that are more conclusive, but also improves the communication between DoD members and leadership. Survey respondents are more inclined to participate in surveys when leadership addresses and listens to their feedback. With quick and objective analysis of survey comments, leadership can address concerns faster and make better-informed decisions.

For the PE survey, this comment analysis approach provides quantifiable results for the reasons F/A-18 and T-45 aircrews and maintainers believe there was an increase in PE within Naval Aviation. The results provide information that help focus the investigation into the plausible root cause(s) of PE. In the DEOCS, the command has a more precise understanding of the types of programs, changes or considerations they need to implement to reduce their command's stress levels to improve overall climate and morale. The analysis from this methodology not only provides quantifiable results, but gathers quality information in a shorter period of time compared to reading all the comments.

B. FUTURE WORK

This research provides additional levels of support to the comment analysis methodology of Cairoli (2017) by further generalizing the method, adding more automation and adjusting the algorithm to provide multiple, non-consecutive labels. Other avenues of future work will enhance the methodology further and improve usability.

1. Other Types of Surveys and Comments

The methodology is generalizable beyond the Navy survey topics in this thesis. This research focuses mainly on survey questions that solicit multiple answers and analyzes surveys in data sets of about 1,000 comments. Further review is required for other types of surveys, particularly surveys where the number of responses is even smaller. For these types of surveys, a reference corpus may be necessary to fill in any gaps in the language or technical terminology.

2. Sentiment Analysis

Opinion-based comments that reflect attitude or behavior are also important and there is opportunity for adding an element of sentiment analysis to our methodology. Moreover, including a component in the comment analysis methodology that also discovers unique comments that are overlooked using the current methodology would be beneficial to highlight potential areas of concern or discover destructive behaviors for leadership to address promptly.

3. Interactive Application

An interactive application (app) that applies the comment analysis methodology would be invaluable. Such an app would allow any Navy service member to more rapidly analyze subsets of survey comments with efficiency and ease. The app could be developed using the R package **shiny** (Chang, Cheng, Allaire, Xie, & McPherson, 2018), which allows analysts to build interactive web apps directly from the R interface. An interactive method with options to include multiple or single labels per comment, filter comments based on demographics and contain other optional features could aid in facilitating more exploration and more rapid analysis of the survey data. The app could also include the option to incorporate and process a new reference corpus or select a specific existing preprocessed reference corpus and even allow the user to supplement the algorithm with acronyms and other jargon. In addition, with such an app there is a need to expand the types of visualizations that might aid the analyst. For example, a word tree, such as those in Figure 12 and Figure 13, could be useful for reading common keywords within the context of several comments.

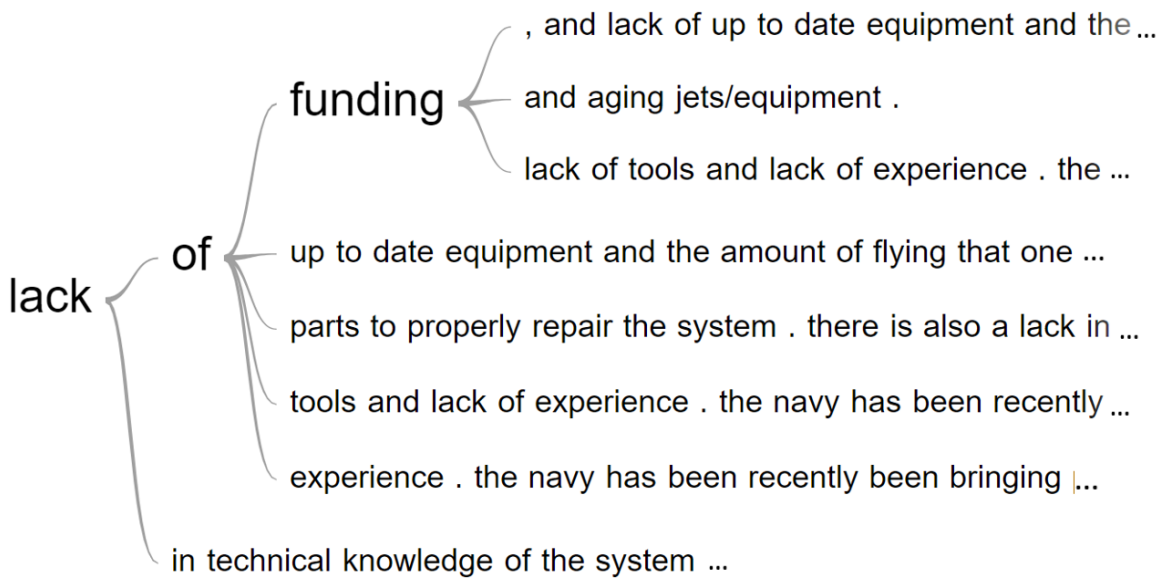

Figure 12. Word Tree

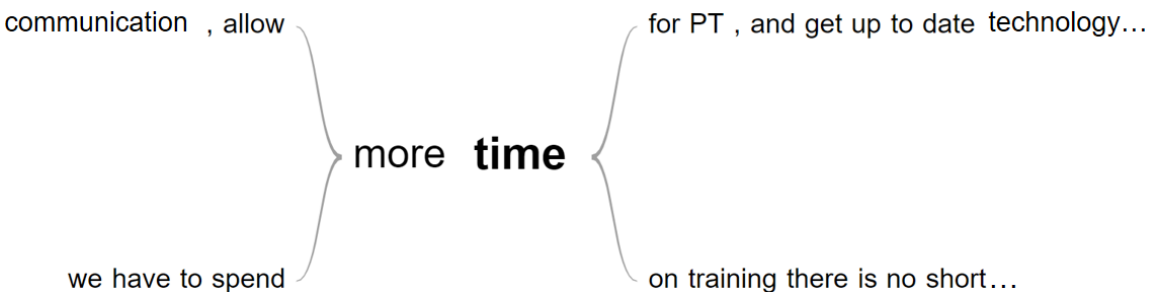

Figure 13. Double Word Tree

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

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

Butts, C. (2015). network: Classes for Relational Data. The Statnet Project. R package version 1.13.0.1. Retrieved from http://CRAN.R-project.org/package=network

Cairoli, C. M. (2017). *Categorization of survey text utilizing natural language processing and demographic filtering.* (Master's thesis). Retrieved from 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). Retrieved from 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. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. https://doi.org/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. https://doi.org/10.1145/2362364.2362367

Csardi G. & Nepusz, T. (2006). The igraph software package for complex network research, *InterJournal*, *Complex Systems* 1695. Retrieved from. http://igraph.org

Department of the Navy. (2005, December 29). *Mission and functions of the Naval Inspector General* (SECNAV Instruction 5430.57G). Washington, DC: Office of the Secretary of the Navy.

Department of the Navy. (2014, May 13). *Naval Aviation Safety Management System.* (OPNAV Instruction 3750.6S). Washington, DC: Office of the Chief of Naval Operations.

Department of the Navy. (2017, July 24). *Navy Equal Opportunity Program* (OPNAV Instruction 5354.1G). Washington, DC: Office of the Chief of Naval Operations.

Feinerer, I., & Hornik, K. (2017). tm: Text Mining Package. Retrieved from https://CRAN.R-project.org/package=tm

Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

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

Physiological episodes within Naval Aviation: Statement before the Tactical Air and Land Forces Subcommittee of the House Armed Services Committee, 115th Cong. (2018) (testimony of Rear Admiral Sara Joyner, Physiological Episode Action Team Lead).

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from https://www.R-project.org/

Silge, J. & Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *Journal of Open Source Software, 1*(3), 37. http://dx.doi.org/10.21105/joss.00037

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

Weiss, S., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining* (2nd ed.) London: Springer London. https://doi.org/10.1007/978-1-4471-6750-1

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

Zhai, C., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining.* New York, NY: ACM Books.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California