



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**APPLYING NATURAL LANGUAGE PROCESSING (NLP)
TO ASSESS HEALTH OF MARINE CORPS CULTURE**

by

Yvonne Sanchez Garcia

September 2022

Thesis Advisor:

Co-Advisor:

Armon C. Barton

Kathryn J. Aten

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
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 2022	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE APPLYING NATURAL LANGUAGE PROCESSING (NLP) TO ASSESS HEALTH OF MARINE CORPS CULTURE			5. FUNDING NUMBERS	
6. AUTHOR(S) Yvonne Sanchez Garcia				
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) The Marine Corps' 38th Commandant calls for a force that can "affect the behavior of real-world pacing threats" through its "innovation, ingenuity, and willingness to continually adapt to and initiate changes in the operating environment." Having a healthy and inclusive environment is vital to the Marine Corps' success in developing and maintaining a force that can rise to the higher standard. Currently, the closest snapshot of the health of Marine Corps culture is collected via command climate surveys, which contain quantitative and free-response questions. Senior levels of Marine Corps leadership do not gain value from the free-text data because 1) analysis on free-text comments is time consuming and no analysis is done on those comments for the commander, and 2) there is no trend analysis on the aggregate of surveys over time and throughout commands, so they are not commander specific. This thesis developed a tool to gain value from free-text comments; it applied natural language processing (NLP) techniques, specifically topic extraction and sentiment analysis, on free-text comments from command climate surveys to provide a quantitative analysis on the health of culture in Marine Corps units. This approach can also be replicated to other services and can inspire efforts that will positively impact talent management and retention.				
14. SUBJECT TERMS natural language processing, NLP, culture, climate, command climate, DEOCS, survey comments, Marine Corps, Marines, text classification, topic extraction, topic modeling, topic discovery, Doc2Vec, KMeans, clustering, CNN, convolutional neural networks, sentiment analysis, Flair			15. NUMBER OF PAGES 73	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**APPLYING NATURAL LANGUAGE PROCESSING (NLP) TO ASSESS
HEALTH OF MARINE CORPS CULTURE**

Yvonne Sanchez Garcia
Captain, United States Marine Corps
BA, University of California, Berkeley, 2016

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

**NAVAL POSTGRADUATE SCHOOL
September 2022**

Approved by: Armon C. Barton
Advisor

Kathryn J. Aten
Co-Advisor

Gurminder Singh
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Marine Corps' 38th Commandant calls for a force that can “affect the behavior of real-world pacing threats” through its “innovation, ingenuity, and willingness to continually adapt to and initiate changes in the operating environment.” Having a healthy and inclusive environment is vital to the Marine Corps' success in developing and maintaining a force that can rise to the higher standard. Currently, the closest snapshot of the health of Marine Corps culture is collected via command climate surveys, which contain quantitative and free-response questions. Senior levels of Marine Corps leadership do not gain value from the free-text data because 1) analysis on free-text comments is time consuming and no analysis is done on those comments for the commander, and 2) there is no trend analysis on the aggregate of surveys over time and throughout commands, so they are not commander specific. This thesis developed a tool to gain value from free-text comments; it applied natural language processing (NLP) techniques, specifically topic extraction and sentiment analysis, on free-text comments from command climate surveys to provide a quantitative analysis on the health of culture in Marine Corps units. This approach can also be replicated to other services and can inspire efforts that will positively impact talent management and retention.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	PURPOSE AND MOTIVATION	1
B.	RESEARCH QUESTIONS.....	2
C.	METHODS	2
D.	BENEFITS OF STUDY.....	2
E.	SCOPE OF THESIS	3
F.	THESIS ORGANIZATION.....	3
II.	BACKGROUND	5
A.	TEXT ANALYSIS AND CULTURE	5
1.	Potential and Value.....	5
2.	Related Work	5
B.	TOPIC EXTRACTION.....	6
1.	Concepts.....	6
2.	Related Work	8
C.	SENTIMENT ANALYSIS	10
1.	Concepts.....	10
2.	Related Work	11
D.	CHAPTER SUMMARY.....	14
III.	METHODS	15
A.	DATA SOURCE.....	15
B.	TOPIC EXTRACTION.....	15
1.	Preprocessing the Data	16
2.	Grouping Data by Question Category	16
3.	Labeling Data with Semi-supervised Approach.....	19
4.	Classifiers.....	20
C.	SENTIMENT ANALYSIS	21
1.	Flair Model Analysis.....	22
2.	Sentiment Statistics.....	22
D.	CHAPTER SUMMARY.....	22
IV.	RESULTS	25
A.	TOPIC EXTRACTION.....	25
1.	Topics	25
2.	Classification Models.....	26
B.	FINAL OUTPUT	27

1.	Question Categories Only.....	27
2.	Topic Labels Only	31
3.	Question Categories and Topic Labels	34
C.	CHAPTER SUMMARY.....	36
V.	DISCUSSION	39
A.	TAKEAWAY.....	39
B.	LIMITATIONS AND CHALLENGES	39
C.	RECOMMENDATIONS.....	40
D.	FUTURE WORK.....	42
E.	CHAPTER SUMMARY.....	43
VI.	CONCLUSION	45
	LIST OF REFERENCES	47
	INITIAL DISTRIBUTION LIST	51

LIST OF FIGURES

Figure 1.	Number of responses per question category	18
Figure 2.	Diagram of pipeline.	23
Figure 3.	CNN training history.....	27
Figure 4.	Sentiments within each question category (numbers only)	29
Figure 5.	Sentiments within each question category (percents only).....	30
Figure 6.	Topic sizes	31
Figure 7.	Sentiments within each topic (numbers only).....	32
Figure 8.	Sentiments within each topic (percents only)	33
Figure 9.	Sentiment for each topic within each question category (percents only) ..	36

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1.	Categories of questions (17 total, labeled zero to 16).....	17
Table 2.	CNN architecture	21
Table 3.	Topic labels (20 total)	25

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
CAOCL	Center of Advanced Operational Culture Learning
CGRI	Commanding General's Readiness Inspector
CNN	Convolutional Neural Network
DEOCS	DEOMI Organizational Climate Survey
DEOMI	Defense Equal Opportunity Management Institute
Doc2Vec	Document to Vector
DTM	Document Term Matrix
EMLS	Exit Milestone Longitudinal Survey
EO	Equal Opportunity
FSMAO	Field Supply and Maintenance Analysis Office
GUI	Graphical User Interface
H&HS	Headquarters and Headquarters Squadron
H&S	Headquarters and Support
IRB	Institutional Review Board
JST	Joint Sentiment/Topic
LDA	Latent Dirichlet Allocation
LDAvis	LDA visual tools
MCAS	Marine Corps Air Station
MCI-W	Marine Corps Installation West
MWCS	Marine Wing Communications Squadron
NaN	Not a Number
NLP	Natural Language Processing

OPA	Office of People Analytics
POS	Part of Speech
PV-DM	Distributed Memory model of Paragraph Vectors
ReLU	Rectified Linear Unit
S&ES	Security and Emergency Services
SAPR	Sexual Assault Prevention and Response
SVM	Support Vector Machine
VADER	Valence Aware Dictionary for sEntiment Reasoner

EXECUTIVE SUMMARY

In the Marine Corps' 38th Commandant's Planning Guidance, General David H. Berger calls for a force that can "affect the behavior of real-world pacing threats" through its "innovation, ingenuity, and willingness to continually adapt to and initiate changes in the operating environment" [1]. Having a healthy and inclusive environment is vital to the Marine Corps' success in developing and maintaining a force that can rise to the higher standards. Currently, the closest snapshot of the health of culture in the Marine Corps is collected via command climate surveys, otherwise called Defense Equal Opportunity Management Institute (DEOMI) Organizational Climate Survey (DEOCS), which is administered by the Office of People Analytics (OPA) [2]. The format of DEOCS survey contains quantitative questions which Marines answer on a numbered scale and free text questions where Marines input any comments they have.

Nearly every Marine Corps unit takes a command climate survey approximately once a year [3]. From available research, it appears that commanding officers do not gain value from the free text responses because they receive these responses in their raw form, without any analysis or highlights of trends in the comment data. Analysis of free text comments is time consuming for the OPA staff that creates the survey reports, and thus no analysis is done on these comments for the command leadership. Similarly, senior levels of Marine Corps leadership do not gain value from the free text responses either because there is no trend analysis on the aggregate of surveys, so they are not commander specific. Commanding Officers and senior levels of Marine Corps leadership, however, are responsible for a multitude of competing priorities and have limited bandwidth and technical expertise to decipher an extensive amount of data.

The Marine Corps needs a streamlined way to extract insights from survey comments at the commander level. This thesis addressed that issue by creating a tool that helps leaders gain value from survey comments, and by extension, a means to assess the health of culture in Marine Corps units.

This thesis focused on analyzing command climate free text responses from the following Marine Corps commands: Security and Emergency Services Battalion, Camp Pendleton, Headquarters and Support Battalion, Camp Pendleton, Headquarters and Headquarters Squadron, Camp Pendleton, H&HS, Marine Corps Air Station Miramar, Marine Corps Installation West headquarters, Marine Wing Communications Squadron 38. The timeframe for the survey data is from July 2020–November 2021.

This thesis developed and applied topic extraction and sentiment analysis to extract insights from survey comments. For topic extraction, the following steps were taken for the comment data: 1) pre-processing; 2) grouping data by question categories; 3) labeling data with a semi-supervised approach of Doc2Vec, K-means clustering, and manual analysis; and 4) building and training two classifiers: Random Forest and Convolutional Neural Networks. For sentiment analysis, a pre-trained Artificial Intelligence model called Flair was used to output a positive or negative label (along with a confidence score); a threshold was established for the confidence level to also classify neutral sentiments.

The question categories that contained the most comment responses were Unit Climate, Personal Experiences and Perception, and Leadership. The four topics that occurred most frequently from topic extraction were Troop Welfare, Leadership, Professionalism, and Morale. The classifier models achieved 89–90% accuracy, which is significant given the unstructured nature of free text. The results provided a sentiment breakdown of question categories only, topics only, and the sentiments for topics specifically found within the question categories.

These detailed reviews arm leaders with insight into the comment dataset, specifically with knowledge of which areas need the most attention. Leaders can then better understand issues by knowing which comments to access for context and can prioritize and focus their efforts on the areas that receive more negative than positive sentiments. Ultimately, leaders are empowered to take action on areas that are affecting servicemembers the most, thereby increasing trust between servicemembers and their leadership and also causing a second order effect of improving troop welfare, readiness, and unit cohesion. Additionally, this approach can be replicated to other services and can inspire efforts that will also positively impact talent management and retention.

References

- [1] Headquarters, U.S. Marine Corps and D. H. Berger, “Commandant’s planning guidance: 38th Commandant of the Marine Corps,” 2019 [Online]. Available: <https://www.marines.mil/News/Publications/MCPEL/Electronic-Library-Display/Article/1907265/38th-commandants-planning-guidance-cpg/>
- [2] Office of People Analytics, “Assessment to solutions,” Defense Equal Opportunity Management Institute [Online]. Available: <https://www.defenseculture.mil/Assessment-to-Solutions/A2S-Home/>.
- [3] Headquarters, U.S. Marine Corps, “Marine Corps Order 5354.1F Marine Corps Prohibited Activities and Conduct (PAC) prevention and response policy,” 2021 [Online]. Available: <https://www.marines.mil/News/Publications/MCPEL/Electronic-Library-Display/Article/1561150/mco-53541f-cancels-mco-53541e-wadmin-ch/>

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

Thank you to my husband, Alejandro. You were most encouraging, and I could always count on your insight and perspective on how I can improve my work. Thank you to my graduating cohort whose sense of humor and willingness to help each other made all the difference in the past two years of CS. I hope to see y'all again in the future.

Thank you to my advisors, Dr. Armon Barton and Dr. Kathryn Aten. You were gracious, patient, and always provided the best guidance and support on an interdisciplinary topic that was new to all of us.

Thank you to Christine Cairoli for pioneering this research in 2017, for providing the continuity for many NPS students after you who did similar research, and for always being willing to help.

A special thank you to the Marines who made this thesis possible—BGen Jason Woodworth, Col John Black, Col Daniel Whitley, LtCol Christine Houser, LtCol Frank Savarese, LtCol Atiim Phillips, LtCol Pedro Ortiz, Col Randy Pugh, and Col (Ret) Todd Lyons. Thank you for believing in my idea and the value this can bring to all units in the Marine Corps.

To LtCol Emmy Hill, thank you for being a phenomenal example and leader; you pushed each of us to do more than we thought we could, you always looked out for opportunities to help us with our goals, and you consistently brought positive energy we could rely on.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

In the Marine Corps' 38th Commandant's Planning Guidance, General David H. Berger calls for a force that can "affect the behavior of real-world pacing threats" through its "innovation, ingenuity, and willingness to continually adapt to and initiate changes in the operating environment" [1]. Having a healthy and inclusive environment is vital to the Marine Corps' success in developing and maintaining a force that can rise to the higher standards.

Additionally, in 2018, a qualitative research project was published by the Center of Advanced Operational Culture Learning (CAOCL) on "Marines' Perspectives on Various Aspects of Marine Corps Organizational Culture" [2]. The researchers identified various cultural issues where a key takeaway highlighted how "simple or crisis-mode solutions are not the right approach" and called the Marine Corps to "invest in long term, deliberative strategies and assessments." Along with implementing long term strategies, however, there needs to be a way to measure where the Marine Corps is today, culturally speaking, and whether the strategies that get implemented are effective or not.

A. PURPOSE AND MOTIVATION

Currently, the closest snapshot of the health of culture in the Marine Corps is collected via command climate surveys, otherwise called Defense Equal Opportunity Management Institute (DEOMI) Organizational Climate Survey (DEOCS), which is administered by the Office of People Analytics (OPA) [3]. The format of DEOCS survey contains quantitative questions that Marines answer on a numbered scale and free text questions where Marines input any comments they have.

Nearly every Marine Corps unit takes a command climate survey approximately once a year [4]. From available research, it appears that commanding officers do not gain value from the free text responses because they receive these responses in their raw form, without any analysis or highlights of trends in the comment data. Analysis of free text comments is time consuming for the OPA staff that creates the survey reports, and thus no analysis is done on these comments for the command leadership. Similarly, senior levels

of Marine Corps leadership do not gain value from the free text responses either because there is no trend analysis on the aggregate of surveys, so they are not commander specific. Commanding Officers and senior levels of Marine Corps leadership, however, are responsible for a multitude of competing priorities and have limited bandwidth and technical expertise to decipher an extensive amount of data.

The Marine Corps needs a streamlined way to extract insights from survey comments at the commander level. This thesis addressed that issue by creating a tool that can provide leaders with a quantitative analysis from survey comments, and by extension, a means to assess the health of culture in Marine Corps units.

B. RESEARCH QUESTIONS

1. What insights can be extracted from free text comments on command climate surveys?
2. How can natural language processing (NLP) techniques be used to extract insights from those free text comments?

C. METHODS

This thesis developed an NLP tool that includes topic extraction and sentiment analysis methods and applied it to free text comments from Marine Corps command climate surveys.

D. BENEFITS OF STUDY

Providing a tool that can analyze and highlight key issues from survey comments equips commanders with the insight and perspective that the quantitative questions simply cannot provide. These insights enable commanders to know which areas need the most attention, to better understand issues by accessing the context behind them, and to be able to prioritize and focus their efforts on the areas that receive more negative than positive sentiments.

By equipping commanders with that knowledge and direction, commanders are empowered to take action on areas that are affecting servicemembers the most, thereby

increasing trust between servicemembers and their leadership, and also causing a second order effect of improving troop welfare, readiness, and unit cohesion.

Given the current issues and focus on culture, retention, and talent management [5], anything that can help leaders correct grievances in their unit is a critical step forward toward improving retention of high caliber members in the service.

Additionally, this approach can be replicated to other services and can inspire efforts that will also positively impact talent management and retention.

E. SCOPE OF THESIS

This thesis focused on analyzing command climate free text responses from the following Marine Corps commands: Security and Emergency Services (S&ES) Battalion, Camp Pendleton, Headquarters and Support (H&S) Battalion, Camp Pendleton, Headquarters and Headquarters Squadron (H&HS), Camp Pendleton, H&HS, Marine Corps Air Station (MCAS) Miramar, Marine Corps Installation West (MCI-W) headquarters, Marine Wing Communications Squadron (MWCS) 38. The timeframe for the survey data is from July 2020–November 2021.

F. THESIS ORGANIZATION

This thesis is organized into six chapters. Chapter II contains background on NLP and various approaches used for topic extraction and sentiment analysis. Chapter III outlines the methods for the pipeline used to output insights from free text comments data. Chapter IV includes the results of inputting the raw comment data into the pipeline. Chapter V summarizes takeaways, limitations and challenges, recommendations, and future work. Chapter VI covers conclusions.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

This chapter discusses the role of text analysis, or text classification, with culture, as well as previous NLP work for extracting insights from free text data through topic extraction and sentiment analysis. This chapter also assesses the applications of NLP techniques in terms of performance, feasibility and scalability, and the value that the research offers future researchers and customers.

A. TEXT ANALYSIS AND CULTURE

1. Potential and Value

As noted in a recent article in the *American Psychologist* journal, “Text analysis can be used to shed light on people and culture,” and help to understand how both culture and people change over time [6]. There are many applications for text analysis, or NLP, from workplace safety to national security, employee retention, product and/or service improvements, to mental health and wellbeing. Examples include using NLP to identify toxic or harassing language from electronic communications before it gets out of hand -- CommSafe AI [7] accomplishes this with NLP; identify insider threats through similar workplace communication analysis; identify creations of extremist groups through propaganda analysis; improve products and services through client feedback analysis; understand quality of life and job satisfaction through survey analysis; and assist with proactive mental health field recommendations.

2. Related Work

In the *American Psychologist* journal article mentioned above, the authors highlighted the power of language at the individual and group level, and how it serves a dual role of providing insight about the creator’s views, attitudes, and potentially emotional and mental states, and the impact that such language has on the receivers [6]. Because of the importance that language plays in the world, specialized tools are needed to help unlock the insights contained in digital text data, which is where topic modeling and sentiment analysis enter the picture. The authors broke down several text analysis methods and the

capabilities that each offer, such as dictionaries to have a baseline lexicon of words, topic modeling to identify central themes, embeddings to capture semantic closeness of words within their context, and neural nets to achieve, according to the authors, “State-of-the-art results [that approach] human level performance” while reducing the manual labor normally required. The authors also provided a diverse range of applications for text analysis, such as dictionaries being used to examine how a category of people speak about specific topics to understand underlying factors (e.g., examine “how employees talk about affiliation or achievement to understand what drives career aspirations and success”); using topic modeling to capture themes in a form of communication and see which themes are more effective (e.g., “Use topic modeling to understand main themes used in persuasive speech”); using embeddings to understand the connection and relations between concepts and perceptions (e.g., “Use embedding to capture perceptions of risk and judgments of meaning, relatedness, and probability”). Even more significant, the authors also noted that word associations can uncover bias for people that may not “make explicitly racist statements” or who are depressed but do not talk about it directly. This research demonstrates how digital text data is rich with insights, and text analysis has the potential to mine those insights and help leaders and organizations be more proactive.

B. TOPIC EXTRACTION

Topic extraction, sometimes referred to as topic detection or topic modeling, has been used to classify and organize large quantities of data. This section reviews relevant concepts and related work to paint a picture on how topic extraction works and how it can be applied.

1. Concepts

a. Stemming versus Lemmatization

Stemming and lemmatization are common NLP techniques meant to remove the inflection of text before conducting further analysis. Stemming, a “rule-based approach” [8], accomplishes inflection removal by chopping off the end of a word and maintaining the stem part of the word; this is a more efficient method that works best when context does not matter.

By comparison, lemmatization, a “dictionary-based approach” [8], accomplishes inflection removal by changing words to their base or dictionary form; this method is more accurate as it takes context into consideration and works best when context is important. For example, with the words “promotion” and “caring,” stemming would return “promot” and “car,” whereas lemmatization would return “promote” and “care.” These approaches are pre-processing options for text classification, depending on the application, stemming or lemmatization can either improve or worsen results.

b. Latent Dirichlet Allocation (LDA)

LDA is a common technique used for topic modeling; it assumes there are latent, or “hidden,” topics in a body of text called a corpus. Topics can be discovered by finding a distribution of probable topics from the words in the corpus and calculating the probability that a word belongs to a certain topic [9].

c. KMeans Clustering

KMeans is an unsupervised machine learning algorithm that clusters numerical data input into a user defined number of clusters based on similarity within the data. Of the many applications, KMeans is used regularly for text clustering [10].

d. Document to Vector (Doc2Vec)

Doc2Vec is a word embedding method that turns documents of any length to the same sized vector representation. This method factors in word order and context, which previous embedding methods have lacked [11].

e. Random Forest

Random Forest is an ensemble method known to address and reduce variance and overfitting issues through its supervised classification. This method is thus commonly used for data sets that have high variance [12].

f. Convolutional Neural Network (CNN)

CNN is a powerful deep learning algorithm that performs well when dealing with high quantities of data that contain a substantial amount of features, such as images. It has also gained popularity with text classification, as unstructured text contains similar characteristics [12].

2. Related Work

a. Military Application

The earliest work on NLP for survey text comments for the military was conducted for the Navy in 2017 by Christine Cairoli [13]. The researchers created a unique approach to have better understanding of why sailors decide to leave or stay in the Navy. They started by applying a label to each comment, then created bins of similar comments. In summary for labeling comments, the researchers 1) preprocessed the data by removing punctuation and stop words, stemming the comments, and replacing common acronyms with a word with similar meaning; 2) identified candidate tokens by constructing a list of unigrams, bigrams and trigrams; 3) calculated a token score for each candidate token based on token specific variables (to include a reference corpus) and comment specific variables; then 4) selected the final label based on the highest scoring candidate token score. For creating topic bins, the researchers created a document term matrix (DTM) of the preprocessed labels, applied LDA to identify the number of topics, identified each topic bin key based on the frequency of topics, then assigned comments to topic bins based on the topic bin key lookup. This method was a trailblazing technique as it integrated demographic data and the researchers were able to home in on low density, high demand roles. Although the end state is invaluable, this approach is unscalable and unlikely to be duplicated in a timely fashion due to topic extraction requiring expert analysts and having a tedious labeling process.

Follow-on research in 2018 took the basis from Cairoli's 2017 research and improved upon it before applying the modified approach to a dataset on the increased occurrence of physiological episodes by T-45 and F/A-18 aircrews [14]. The researchers improved upon the approach by 1) making the approach more generalizable by taking the

comment-label pairs from Cairoli's [13] dataset and using them to estimate regression coefficients for use during label score calculations, 2) removing the "need to acquire and preprocess an external reference corpus," 3) increasing automation for discovering topics by extracting the most frequent, salient, and distinct words to draft an initial topic bin key list, and 4) incorporating multiple topics per survey comment by treating each sentence as a separate document. As an improved approach, this method continued to rely on previously labeled comments as a starting point for regression coefficients and required extensive analyst input to determine the final bin topics. This approach improved Cairoli's [13] approach from multiple angles, however, it remains unscalable and unlikely to be duplicated in a timely fashion as it requires previously labeled comment data and significant analyst involvement.

Follow-on research in 2020 took the basis of both Cairoli [13] and Layug [14] and improved upon it by utilizing 1) lemmatization, instead of stemming, for the initial list of topics, 2) LDA visual tools (LDAvis) and relevancy metrics for bin topics, and 3) by creating a visual application that streamlines many parts of the process via a package called Shiny in R [15]. Lemmatization improved clarity when researchers evaluated candidate labels; LDAvis, an "internet based" model, allowed the user to scale a lambda value up and down to see the effects on how the parameter value can narrow down the list of bin topics to a suitable length of only the most relevant topics. Additionally, the researchers stored the second and third highest scoring candidate labels to assist with accurately assigning comments into topic bins that they most align with, therefore reducing the size of the "other" category for topics. The author took a process that formerly only savvy operational analysts were familiar with and made a helpful visual flow. This approach, which built upon the work of two previous researchers work, is the first of its kind within the Navy that has the potential to become increasingly accessible.

b. Non-Military Application

(1) Doc2Vec and CNN

In 2021, researchers from Turkey utilized Doc2Vec methods to embed both Turkish and English news data sets and used several classifiers to measure effectiveness [12]. From

the original news data set, they created four versions: original, cleaned data, stemmed data, and cleaned and stemmed data. They created a Doc2Vec model for each data set and ran it through a deep learning model, CNN, as well as machine learning algorithms such as Gaussian Naïve Bayes, Random Forest, Naïve Bayes, and Support Vector Machines (SVMs). The CNN model consisted of four convolutional layers, with three pooling layers in between, followed by a flattening layer, and dense layers with the final dense layer using Softmax for classification. Through their comparisons, they achieved best performance with the Turkish data set when using the “Clean + Stem” version with a CNN model utilizing “distributed memory model of paragraph vectors (PV-DM);” the English datasets achieved best results with the “Clean” version with a CNN model utilizing PV-DM. The researchers’ work demonstrates the valuable contributions of combining Doc2Vec and CNN for text classification for topics and labels.

(2) LDA and KMeans

A 2013 approach by Cheng combined LDA and K-means clustering to detect new topics in both a well-curated dataset, such as articles from *New York Times*, and a less structured source such as Facebook comments [16]. The researchers first used LDA to reduce dimensionality and transform word space into topic space, then explored multiple approaches for detecting an “anomaly,” which was a new topic. The approaches included: mixture models, histograms, support vector machines, and clustering; the researchers found that KMeans clustering, an unsupervised method, performed the best. The combination of techniques had not been done previously, and the researchers achieved better results for the New York Times articles than the shorter comments from the Facebook data. This research suggests that the combination of approaches, LDA and KMeans clustering, is not well suited for short text comments that are less topic coherent.

C. SENTIMENT ANALYSIS

1. Concepts

Sentiment analysis, sometimes referred to as sentiment polarity or orientation, has been used in many studies for opinion mining and to enhance text summarization [17]. This

section reviews sentiment analysis approaches used by different researchers and some of the domains it has been applied to.

a. Rule Based Sentiment Analysis

(1) VADER Sentiment

Valence Aware Dictionary for sEntiment Reasoner (VADER) Sentiment is a rule-based sentiment analyzer that output the probability that a comment is positive, neutral, or negative. This analyzer also calculates a compound score that considers all lexical features; this compound score then determines the sentiment label based on a user defined threshold. A significant downside with VADER is that it does not factor in context within in its assessment [18].

(2) TextBlob

TextBlob is another rule-based sentiment analyzer that calculates sentiment polarity and subjectivity for the words it recognizes. This approach outputs a float value, which a user then needs to establish a relevant threshold to determine a label for positive, neutral, or negative sentiment [19].

b. Embedding Based Sentiment Analysis

Flair is an embedding based model that outputs a sentiment label of positive or negative, along with a confidence score for that label [20]. When compared to the rule-based sentiment analyzers, embedding based models provide more flexibility, and thus accuracy, due to the text embedding, making the rule-based sentiment analyzers more rigid. Users are also able to modify the output by thresholding the model's confidence score.

2. Related Work

a. Custom Built Methods

In a 2020 article in the *Applied Clinical Informatics* journal, a hybrid approach was used for topic and sentiment analysis of patient experience feedback [21]. In this article, health care agencies received patient experience survey comments along with associated labels for care categories and sentiments (such as positive, negative, neutral, or mixed)

from a separate agency that collects the surveys. The authors sought to classify the mixed comments; they broke them down to the sentence level, then developed and trained a neural net model with dense and dropout layers to further classify mixed comments into positive, negative, or neutral. By combining the assessment of positive, negative, or neutral and the different labels of care categories, the researchers identified the most frequently occurring words and bigrams to gain context behind the comments' sentiment. Through their work, the authors had a quick snapshot of patient feedback and were able to focus their comment reviews on those care categories that received more negative than positive sentiments (e.g., temperature control was a big issue in the "room" category; long waits to be discharged was the biggest issue in the "discharge" category). The researchers did not pursue their own way to identify care labels nor initially labeling comments as positive, negative, neutral, or mixed. Their work highlights the follow-on actions that can be taken after comments have been categorized into topics and had a sentiment label applied.

In 2013, sentiment analysis was conducted on Twitter posts using machine learning techniques [22]. The authors took Twitter posts that related to user reviews on electronic products, created a feature vector for those posts, and used an ensemble of classifiers to determine the sentiment of each post. For the pre-processing step, the researchers corrected misspellings and replaced slang words, another manual labor point, and manually labeled 1200 posts as positive and negative. The researchers then created a feature vector to overcome the challenge of analyzing posts that are 140 characters or less. The authors created their feature vector in two phases—first, they extracted Twitter specific features, and then conducted further feature extraction on the remaining simple text. For Twitter's specific features, the authors identified the amount of positive and negative hashtags, presence of emoticons and positive or negative value if they were present, and then removed both hashtags and emoticons from each post, leaving only the simple text. For further feature extraction, the researchers created and maintained a negative keyword list, positive keyword list, and a list of different words that represented negation. They then selected a special keyword that was a relevant part of speech (POS) for each post. The authors utilized and compared classifier performance for Naïve Bayes, SVM, Maximum Entropy, and an Ensemble Classifier. The SVM, Maximum Entropy, and Ensemble (voting

rule) classifiers performed well across the board with 87–93% on accuracy, precision, and recall metrics. The Naïve Bayes classifier achieved 88–91% on accuracy, precision, and recall metrics, which is a very competitive performance. The authors attribute the strong classifier performance to the features they selected as part of the feature vector. In total, their feature vector had eight relevant features: POS (if the keyword selected was a relevant POS such as adjectives, adverbs, or verbs), special keyword (if a keyword was selected and if it was positive, negative, or none was selected), presence of negation, presence of emoticon, number of positive keywords, number of negative keywords, number of positive hashtags, and number of negative hashtags. Although this research required a significant amount of manual work for the feature vector, their approach resulted in strong performance and thus provided a framework for how any organization can build a pipeline that is tailored to their domain.

b. Unsupervised Joint Sentiment/Topic Model

In 2009, research on topic extraction and sentiment was conducted and suggested that “sentiment polarities [largely depend] on topics or domains” [17]. The research demonstrated the feasibility of having an unsupervised method that conducts sentiment and topic analysis simultaneously, called Joint Sentiment / Topic (JST) model [17]. Having concurrent sentiment and topic detection is exceptionally helpful for opinion mining and summarizing text. The JST model uses the LDA framework and adds a sentiment layer. As the report noted, the authors “[improved] sentiment detection accuracy” by incorporating minimum prior information along with the JST model; this includes: a paradigm word list, mutual information, full subjectivity lexicon, and a filtered subjectivity lexicon. The JST model was tested on a movie review dataset and focuses on document level analysis but can be extended to any choice of granularity, e.g., sentence, word, or phrase level. The authors also compared the performance of their standard JST model, which has a topic distribution for every document, and a Tying JST model, which only has one topic distribution for all documents. Their models treat each “document as a bag of words,” which inherently ignores the order of words. This prevented statements such as “not good movie” and “not bad movie” from being accurately classified as “bad” and “good,” respectively. The authors’ future work includes simple improvements such as using

bigrams and trigrams to represent “higher order information” in documents, which can drastically impact the reported performance of the JST model with various forms of prior information. This approach is appealing due to the unsupervised nature, which equates to scalability and ease of use; but organizations would need to invest in creating the relevant minimum prior information for their domain to truly make the JST approach effective.

D. CHAPTER SUMMARY

This chapter assessed the value and applications of text analysis to culture through topic extraction and sentiment analysis. Military application of NLP techniques has been limited to topic extraction and only applied to Navy specific surveys for sailor retention, and specific aircrews surrounding stress causes and an increase of physiological episodes. Based on available research, the Naval Postgraduate School researchers were the first to apply any kind of NLP on survey text data from the military.

This review revealed there is limited research on applying both topic and sentiment analysis, and the research is even more limited for working with data of short free text comments, such as those found from survey responses, and even more limited for military free text data. The next chapter discusses the methodology used to extract topics and analyze sentiments on Marine Corps command climate survey data.

III. METHODS

The goal of this thesis is to develop a method that can gain key insights from the free text responses, and by extension, a means to assess the health of culture in Marine Corps units. This chapter describes the pipeline built to achieve that goal, which incorporates the data source background, topic extraction through a semi-supervised approach, and sentiment analysis for each topic. Topic extraction covers the methods used to classify and organize the comment data set into topic. Sentiment analysis covers the pre-trained model used to obtain a sentiment label for each comment.

A. DATA SOURCE

This thesis analyzed free text comment data from 11 command climate surveys from Marine Corps commands: S&ES Battalion Camp Pendleton, H&S Battalion Camp Pendleton, H&HS Camp Pendleton, H&HS MCAS Miramar, MCI-W headquarters, MWCS 38. The timeframe for the survey data was from July 2020–November 2021.

The data used fully complies with the Naval Postgraduate School Institutional Review Board (IRB) standards, as it is pre-collected data. The data itself did not include privately identifiable information of survey participants. The survey comments were stored in separate pdf files with questions and responses for each question; each one was converted into an Excel file, where the headers and footer information were removed. The cleaned-up Excel files were read into a master file using a Pandas dataframe to be ready for preprocessing.

B. TOPIC EXTRACTION

As discussed in Chapter II, topic extraction is a means to classify and organize a large set of data. The steps followed for topic extraction include preparing the data through preprocessing steps, determining the set of topic labels, assigning a label for each comment, and training two model classifiers to predict topics on new data.

1. Preprocessing the Data

The first step in the pipeline includes cleaning and preparing the data for the methods in the follow-on steps. Each comment had the following pre-processing steps done in the order depicted:

1. Comments with Not a Number (NaN) values were removed.
2. Repeated characters were removed.
3. Contractions were expanded.
4. Special characters were removed.
5. Comments with three words or less were removed, as these are too short to effectively communicate a topic or context behind a sentiment.
6. Stop words were removed.
7. Empty comments were removed.
8. A unique identifier was added to the remaining comments.

The listed pre-processing steps are common pre-processing steps for text analysis [24], which are critical for extracting topics and sentiments from the comment data. The resulting number of comments after cleaning were 4,987.

2. Grouping Data by Question Category

The dataset covers multiple surveys from different timeframes, and because of that, there are variations of the core questions being asked. To account for the variation in questions from multiple surveys, the questions were grouped into categories that had a similar core question. Table 1 shows the final category list of 17 groups and Figure 1 shows the respective size (number of comments) of each category group. Of note, the three question categories that contained the most comment responses were Unit Climate, Personal Experiences and Perception, and Leadership. This kind of summary indicates to leaders the categories that elicited the most responses, and thus which areas their Marines are most interested in or concerned with.

Table 1. Categories of questions (17 total, labeled zero to 16)

CATEGORY #	QUESTION CATEGORY
0	Equal Opportunity
1	Organizational Effectiveness
2	Sexual Assault Prevention and Response (SAPR)
3	Unit Cohesion
4	Communication
5	Morale
6	Professionalism/Treatment of Others
7	Accountability
8	Personal Education
9	Retention
10	Stress
11	Quality of Life / Barracks
12	Unit Climate
13	Performance Feedback
14	Leadership
15	Personal Experiences and Perception
16	General Comments

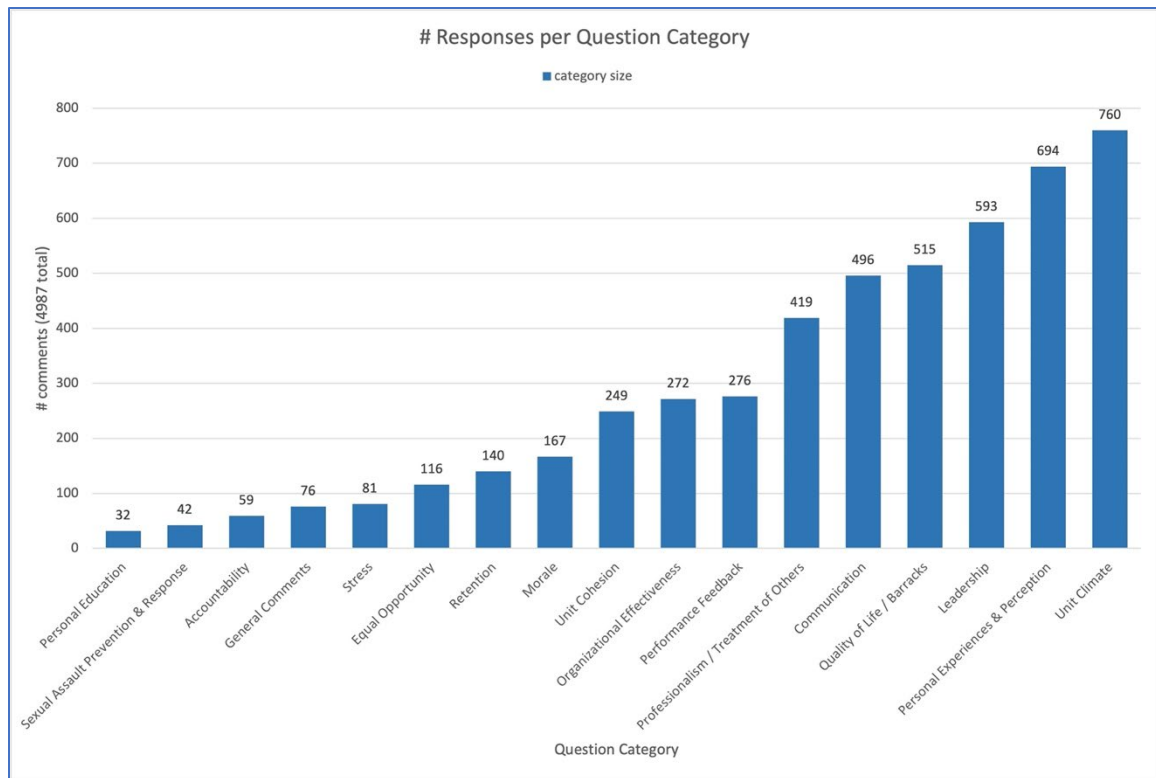


Figure 1. Number of responses per question category

The categories of Professionalism / Treatment of Others and Accountability were intentionally separated from Leadership. Leadership was considered more for actions and responsibilities (or the “what”), whereas Professionalism / Treatment of Others was considered more as the “how” an action gets carried, and “how” someone treated another person; Accountability was seen as a special case of Leadership where someone is or is not held accountable for their actions.

The category of Personal Education, SAPR, and Stress were among the smallest size of groups that had fewer than 100 responses each. Despite their small size, these categories were intentionally chosen because of the emphasis Marine Corps leadership has placed in 1) improving the education of our servicemembers to develop critical thinking, 2) eliminating sexual assault in our forces due to the trust it erodes in a unit, and 3) the severity of servicemember suicides both while on active duty and as veterans.

3. Labeling Data with Semi-supervised Approach

Within each question category, each comment was labeled with a topic using a semi-supervised approach. The following steps outline the process within each category group: creating an embedding layer through Doc2Vec, clustering the data with KMeans clustering, determining a topic label for each cluster after manually assessing the cluster center and immediate neighbors, and applying topic labels to the respective clusters.

a. Doc2Vec

The comment data within each category group was converted into a 50-dimensional vector via an approach from the Python gensim library called Doc2Vec. Of the many approaches for transforming words into an embedding layer, Doc2Vec method was chosen due to its ability to retain context since it vectorizes the comment itself in addition to the individual words in the comments [11].

b. K-means

The vectors within each category group were clustered using K-means clustering from the Python scikit-learn library. Based on the size of the category group however, the groups with fewer than 300 comments were clustered into five clusters, and those with more than 300 comments were clustered into ten clusters.

K-means was used due to its unsupervised nature for grouping samples in the data set according to their similarity in vector space. This allows the overall approach to be duplicated and implemented more easily.

c. Manual Analysis

Within each question category group, the original comments associated with the data point closest to each of the cluster centers, including the five nearest neighbors to the center, were extracted for manual assessment. Each comment covered many topics, and as each comment was assessed, every new topic was written down and received a vote once per comment if it applied. The topic with the highest vote was chosen as that cluster's topic

label. If there were two topics with equally high votes, both topics were chosen as that cluster’s topic label.

If the five immediate center neighbors resulted in a tie amongst more than two topics, then 10 immediate center neighbors were used to break the tie. This was done in an effort to represent the clusters more accurately within each category group.

Additionally, the topic labels repeated amongst different clusters, which reduced the comment data set from the original number of clusters to a manageable number of topics. Step 3c changed the unsupervised approach into a semi-supervised approach.

d. Apply Topic Labels

Once all topic labels were determined, each topic label was applied to its respective cluster.

4. Classifiers

Two classifiers, Random Forest and CNN were chosen to validate the semi-supervised approach and measure how accurate the models can predict topics on new data. Random Forest, a supervised machine learning algorithm, was chosen as one of the classifiers due to its simplicity, efficiency, and ability to minimize overfitting from its ensemble learning. CNN, a deep learning algorithm, was chosen as another classifier due to its ability to extract features from “the simplest to the most complex features” [12]. The CNN model consisted of convolutional layers with rectified linear unit (ReLU) activations, max pooling layers, a flatten layer, a fully connected layer with ReLU activation, and a SoftMax classification layer, as this combination has had previous success [12]. Table 2 shows the layer architecture.

Table 2. CNN architecture

#	LAYER TYPE	SIZE
1	Convolutional + ReLU	Kernel = 3, Filters = 64
2	Max Pooling	Pool size = 2x2
3	Convolutional + ReLU	Kernel = 3, Filters = 64
4	Max Pooling	Pool size = 2x2
5	Convolutional + ReLU	Kernel = 3, Filters = 64
6	Max Pooling	Pool size = 2x2
7	Convolutional + ReLU	Kernel = 3, Filters = 64
8	Flatten	
9	Fully Connected + ReLU	128
10	Softmax	20

The comment data was split into an 80/20 split where 80% of the data was used as the training data, and 20% of it was used for the testing data set, as well as a 70/30 split. The Random Forest model was trained on both the 80/20 split training data and the 70/30 split data, and the CNN model was trained on only the 80/20 split training data for 200 epochs. The accuracy of each model's ability to predict topics on new data was then measured with the test set.

C. SENTIMENT ANALYSIS

As discussed in Chapter II, sentiment analysis is commonly used to extract the sentiment orientation from comments. A pre-trained model, called Flair, was chosen to obtain a sentiment label for each comment, followed by a statistics breakdown of sentiment labels for the question category dataset, the topic dataset, and each topic found within each question category.

1. Flair Model Analysis

Three common pre-trained models for sentiment analysis, VADER, TextBlob, and Flair were explored, and the Flair model was chosen. Through manual observation of several comments, Flair was most consistently accurate, over 95% of the time. Flair was assessed to perform best due to the fact that it is a pre-trained model that utilizes text embeddings, whereas VADER and TextBlob are rule based models (and not machine learning models) [20]; more specifically, VADER does not take context into consideration, and TextBlob only outputs a score between -1 (for negative) and 1 (for positive), and the thresholds for categorizing each negative, neutral, and positive label would require a deeper look [18], [19].

Without modifications, Flair takes in raw text and outputs a value of “Positive” or “Negative” along with a score of how confident the model is in that label. After manual observation of several comments, the Flair approach was modified to also output a “Neutral” value when the confidence level was below a certain threshold. A threshold of 70% was chosen to adequately account for comments that are not strictly positive or negative. Each comment received a sentiment label assessed by Flair.

2. Sentiment Statistics

For the final step in the pipeline, the sentiment analysis statistics were extracted for the question category dataset, the topic dataset, and each topic found within each question category. All three data sets had a breakdown of how many comments were labeled as positive, neutral, and negative.

D. CHAPTER SUMMARY

Figure 2 shows the steps described above along with the final output: a numerical breakdown of positive, neutral, and negative comments for each question category, topic, and topics within each question category.

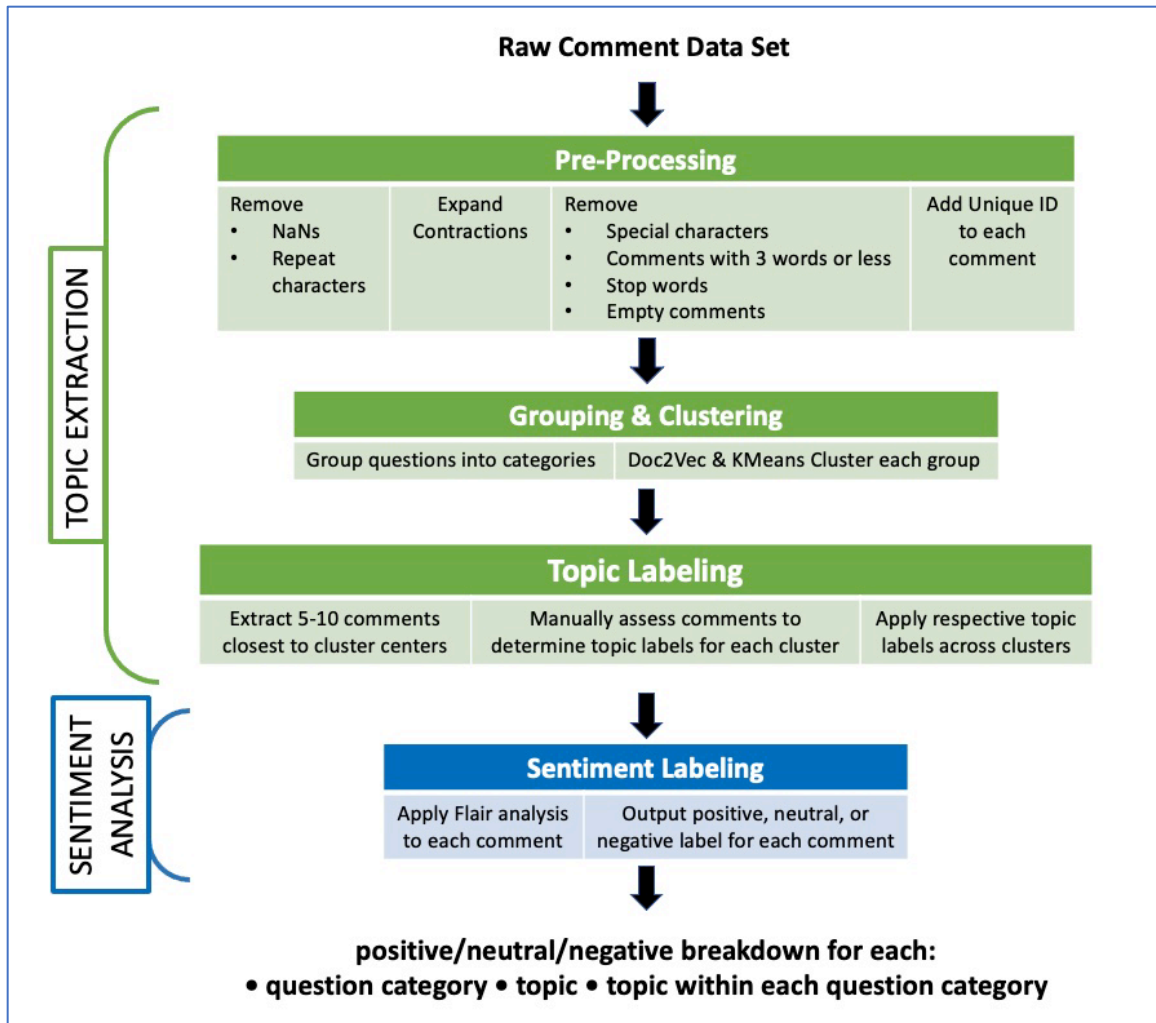


Figure 2. Diagram of pipeline.

This chapter discussed the pipeline used to extract insights from raw command climate survey comments. This kind of assessment would allow leaders to (1) have a quick snapshot of areas that need the most immediate attention, (2) know which areas are going well, and (3) have direct access to the comments pertaining to each topic category and sentiment label to gain more context.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

This section discusses the results from the topic extraction process and the final output for each question category, topic, and topics within each question category.

A. TOPIC EXTRACTION

1. Topics

After manual analysis, there were nine single topic categories, with an additional 11 dual topic categories. Each single and dual topic was treated as a distinct label, and thus the final list of topics had 20 distinct topic labels as summarized by Table 3.

Table 3. Topic labels (20 total)

Topic Labels			
Single Topic		Dual Topic	
1	Accountability	10	Accountability and Leadership
2	Communication	11	Accountability and Professionalism/ Treatment of Others
3	Feedback	12	Feedback and Troop Welfare
4	Leadership	13	Leadership and Communication
5	Morale	14	Leadership and Resources
6	Opportunities	15	Leadership and Troop Welfare
7	Professionalism/ Treatment of Others	16	Morale and Leadership
8	Retention	17	Morale and Professionalism/ Treatment of Others
9	Troop Welfare	18	Morale and Troop Welfare

Topic Labels			
		19	Professionalism/Treatment of Others and Leadership
		20	Troop Welfare and Facilities

The topics of Professionalism / Treatment of Others and Accountability were intentionally separated from Leadership to capture the following nuances:

- Leadership was considered more for actions and responsibilities (the “what”) like decisiveness, setting expectations, policies and procedures, and whether a leader showed they care about their Marines.
- Professionalism / Treatment of Others was considered more as the “how” an action gets carried, and “how” someone treated another person. This includes aspects like fairness, equal consideration in punishment, ego, and behavior characteristics.
- Accountability was seen as a special case of Leadership where someone is or is not held accountable for their actions

Additionally, troop welfare considered aspects like physical health, mental health, barracks living conditions, whether Marines are recognized for their work, and work life balance.

2. Classification Models

For the classification models, the Random Forest model had 89% accuracy for both 80/20 and 70/30 train test split; the CNN model achieved 90% validation accuracy with 42% validation loss as shown in Figure 3. This means they both achieve an 89%–90% accuracy when predicting topic labels for new comment data.

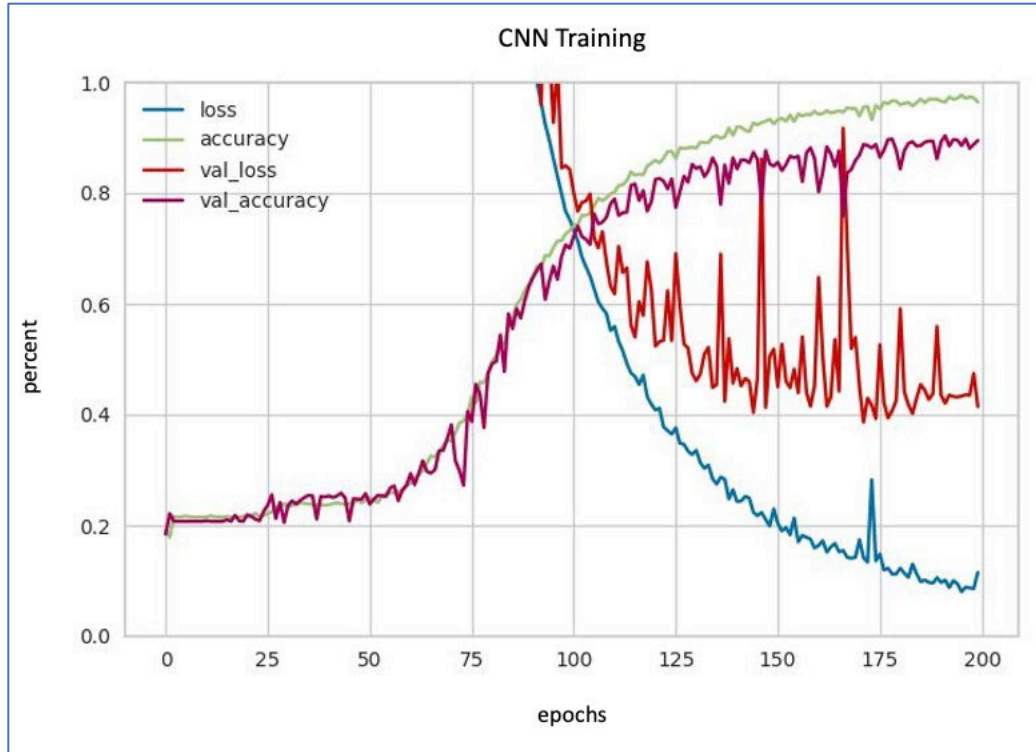


Figure 3. CNN training history

B. FINAL OUTPUT

The final output is a numerical breakdown of sentiments (positive, neutral, and negative) for comments in each question category, a summary of topics found in the overall dataset along with a sentiment breakdown per topic, and a summary of topics found in each question category and the respective sentiment breakdown per topic. This kind of visual of topics and sentiments gives a snapshot overview of the overall comment dataset as well as the respective question categories.

1. Question Categories Only

In Figure 4, the data was sorted by the number of total comments in each question category, and it shows the number of comments that were labeled as positive, neutral, and negative per question category. Overall, Personal Experiences and Perception had more negative comments than any other category, and Unit Climate had more positive comments than any other category. This kind of summary indicates to leaders that those categories

merit special review to understand deeper context of contributing factors and can start by reviewing the topics found within those categories.

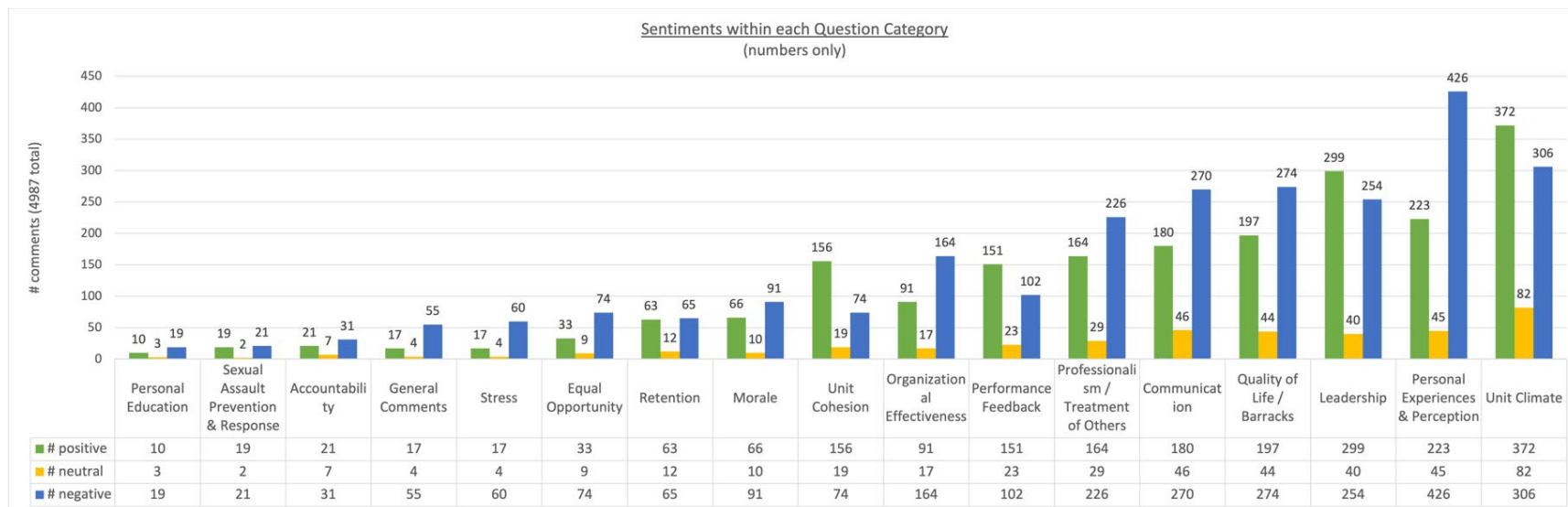


Figure 4. Sentiments within each question category (numbers only)

Figure 5 shows a different perspective from Figure 4; it shows of the percentage of comments labeled as positive, neutral, and negative within each question category. The category with the largest percent of negative comments was under the Stress category, and the one with the largest percent of positive comments was under the Unit Cohesion category. This kind of summary provides leaders another way to assess which category sections merit special review to understand deeper context.

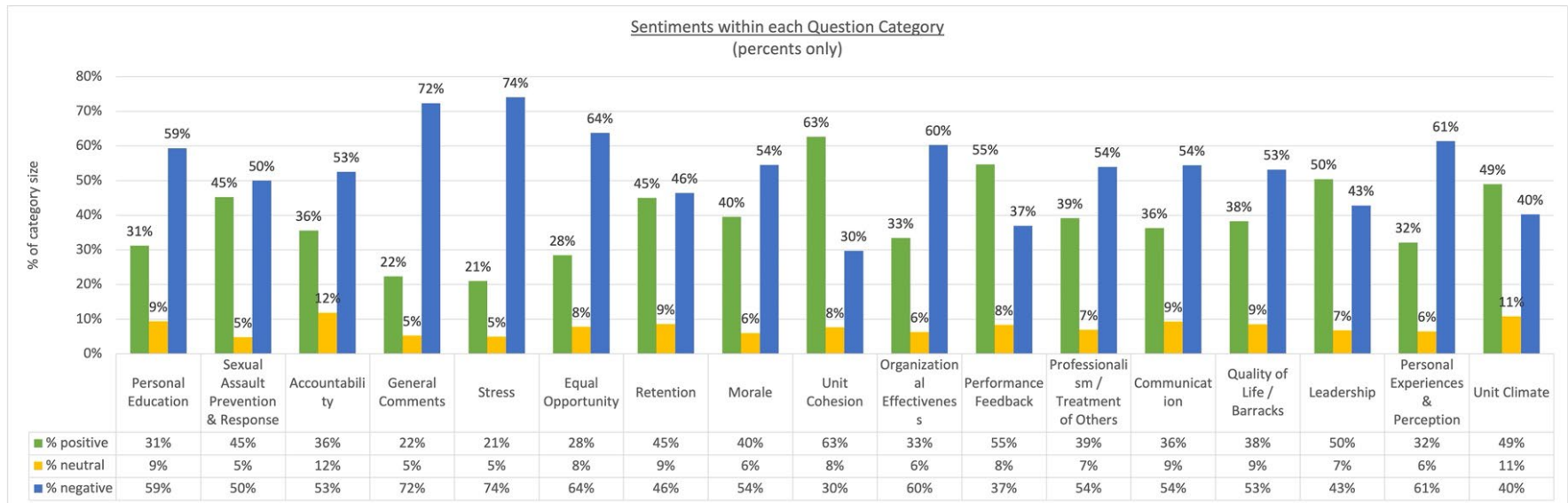


Figure 5. Sentiments within each question category (percents only)

2. Topic Labels Only

Figure 6 shows each topic label and its respective size (number of comments in a topic), and the data was sorted by the number of total comments in each topic. Of note, the four topics that occurred most frequently from topic extraction were Troop Welfare, Leadership, Professionalism/Treatment of Others, and Morale. This kind of summary indicates to leaders the topics that were most discussed by their Marines, and thus the areas their Marines are most interested in or concerned with.

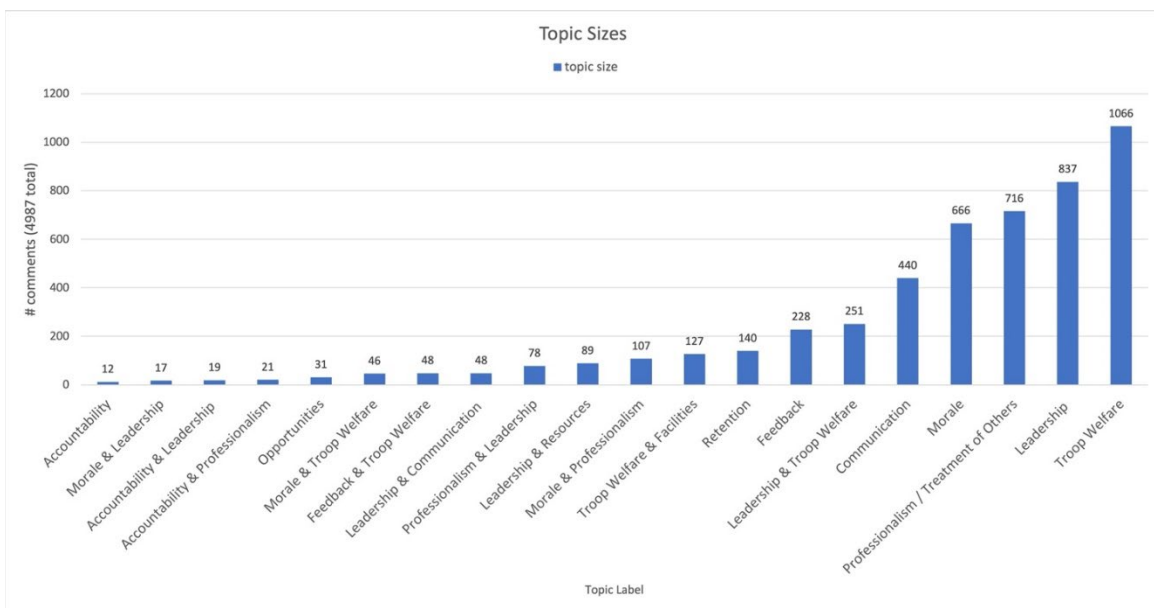


Figure 6. Topic sizes

In Figure 7, the data was sorted by the number of total comments in each topic (topic size), and it shows the number of comments that were labeled as positive, neutral, and negative for each topic. Troop Welfare had more negative comments than any other topic, and more positive comments than any other topic. This kind of summary indicates to leaders that comments relating to Troop Welfare merit special review to understand what is causing or contributing toward the strong negative and positive responses.

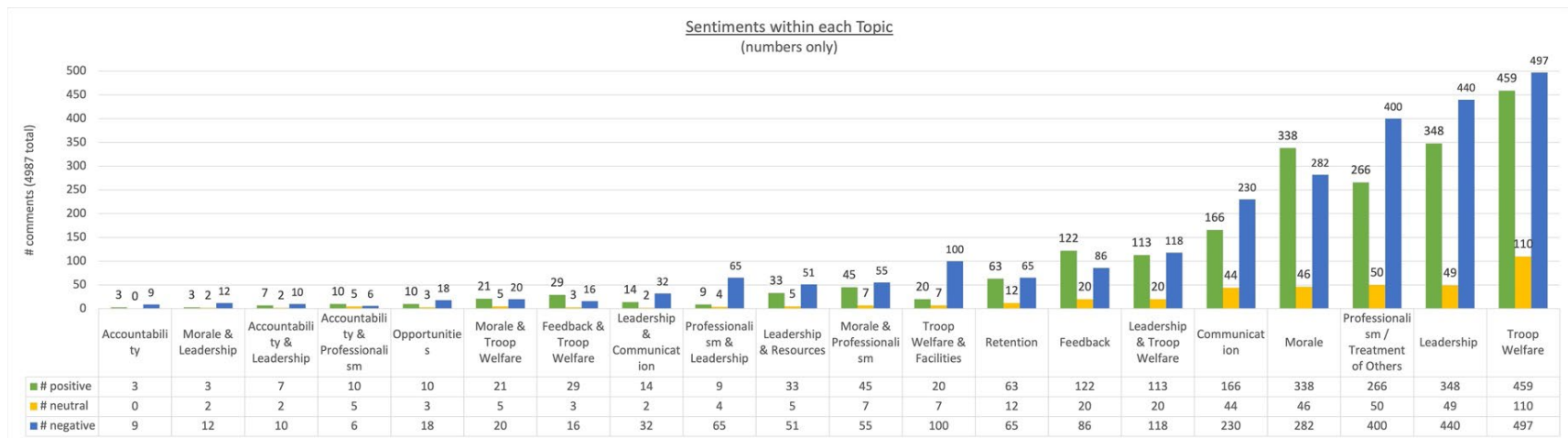


Figure 7. Sentiments within each topic (numbers only)

Similar to the question categories, Figure 8 shows a percentage perspective of the number of comments labeled as positive, neutral, and negative within each topic. The Professionalism / Treatment of Others and Leadership dual topic had the largest percent of negative comments, and the Feedback and Troop Welfare dual topic had the largest percent of positive comments. This kind of summary provides leaders another way to assess which topics merit special review to understand deeper context of contributors.

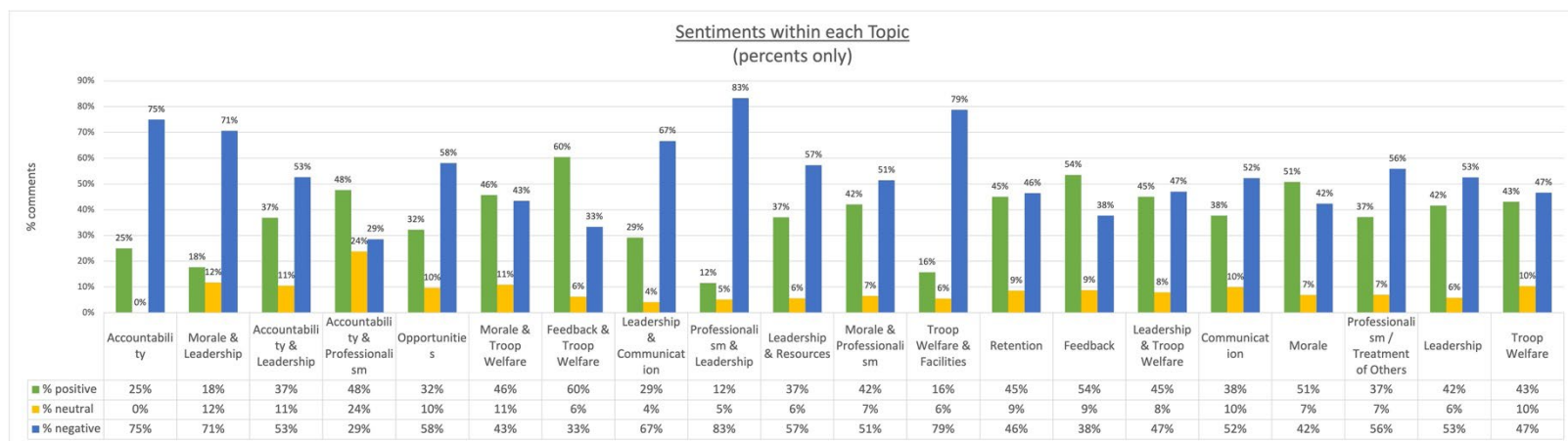
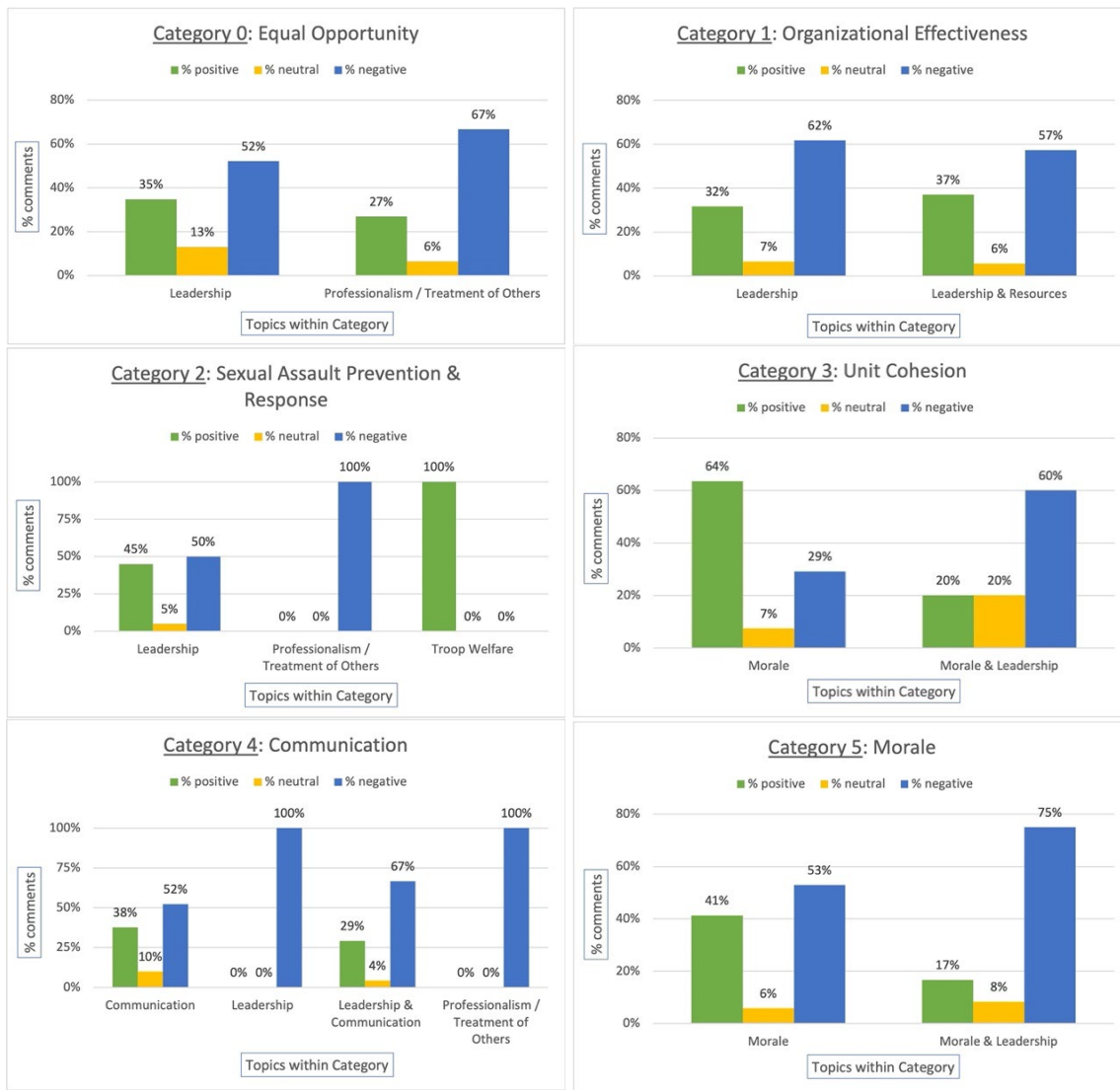
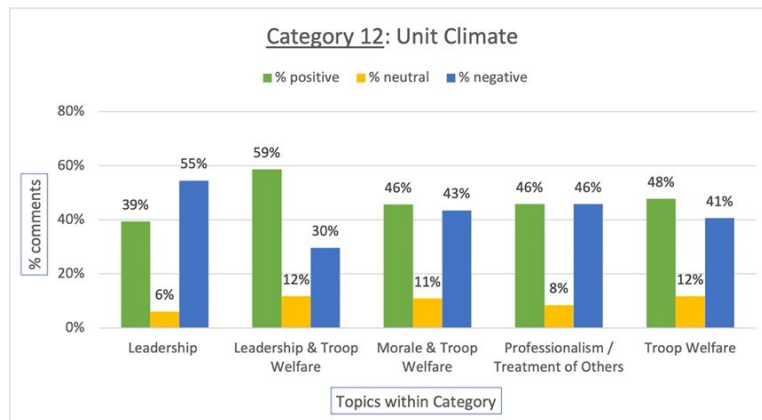
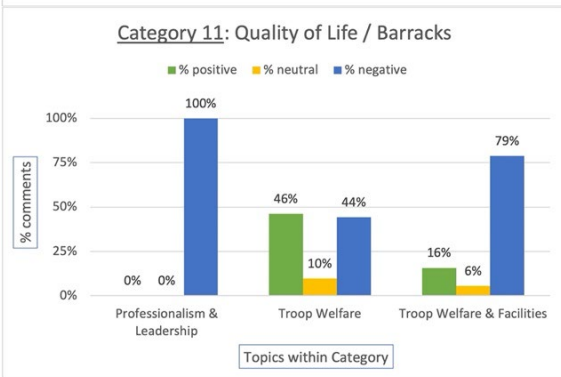
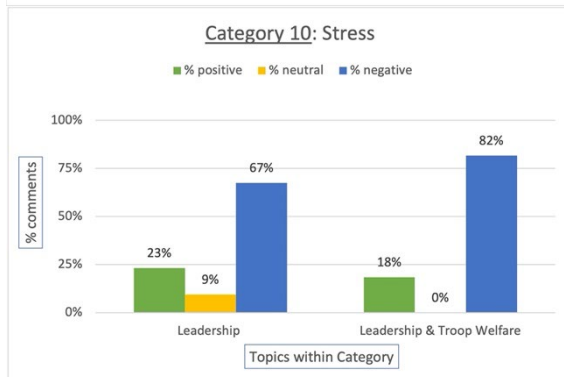
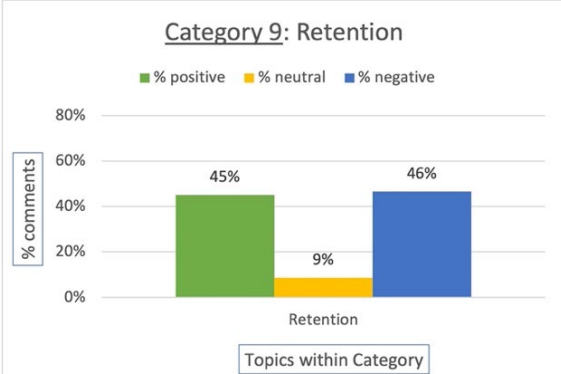
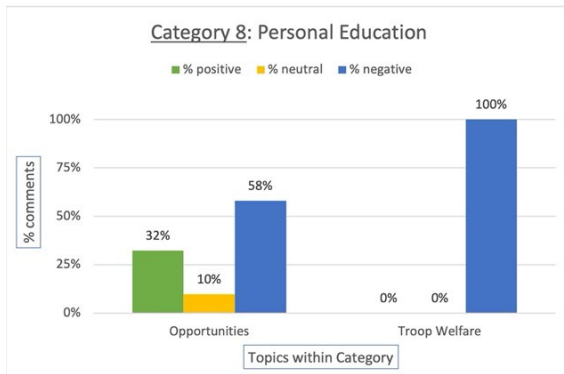
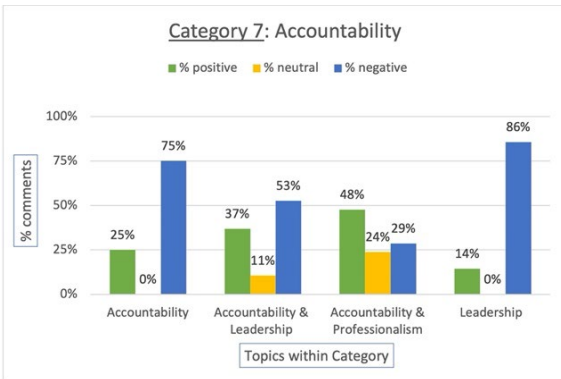
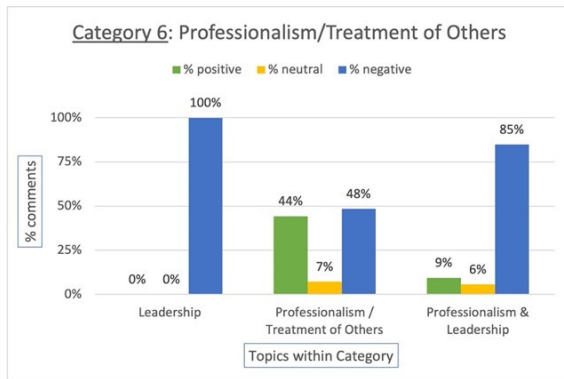


Figure 8. Sentiments within each topic (percents only)

3. Question Categories and Topic Labels

Figure 9 shows a table of figures containing the sentiment breakdown for each topic within each question category. The values are in percentages, and as a result, topics that show 100% of a particular sentiment only had one or two comments. This kind of detailed summary, nonetheless, arms leaders with knowledge of the topics discussed in each question category and provides direction for a focused review on the previously flagged question categories from Figures 4 and 5: Personal Experiences and Perception, Unit Climate, Stress category, and Unit Cohesion.





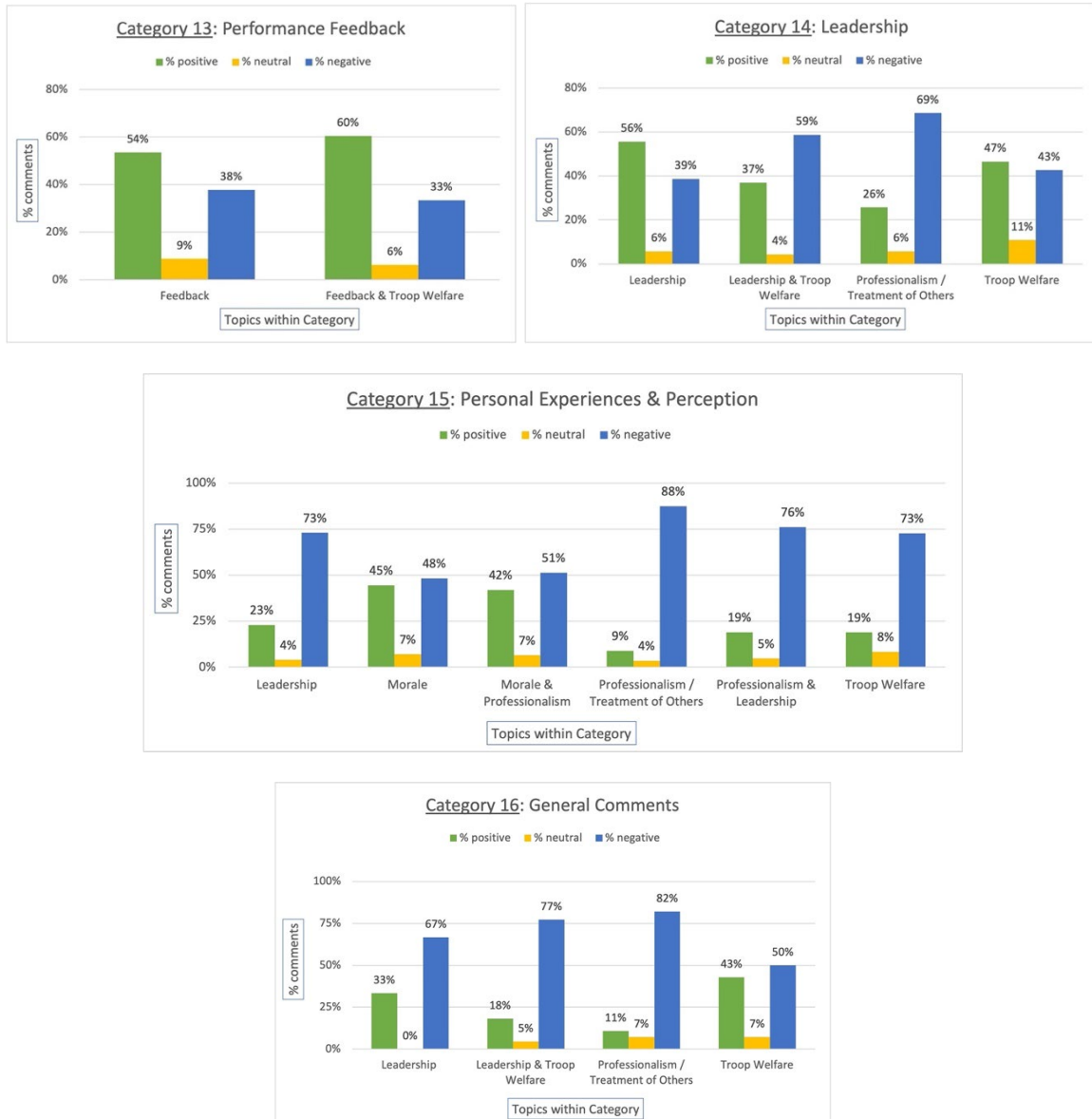


Figure 9. Sentiment for each topic within each question category (percents only)

C. CHAPTER SUMMARY

This chapter discussed the topic labels found during the topic extraction process, the classification model results, and the final output: a numerical breakdown of sentiments (positive, neutral, and negative) for comments in each question category, a summary of topics found in the overall dataset along with a sentiment breakdown per topic, and a

summary of topics found in each question category and the respective sentiment breakdown per topic.

Each visualization of topics and sentiments gave a snapshot overview of the comment dataset by highlighting which topics Marines are most interested in or concerned with and which topics / question categories merit special review to understand contributing factors. In summary, these detailed reviews arm leaders with insight into the comment dataset, specifically with knowledge of which areas need the most attention. Leaders can then better understand issues by knowing which comments to access for context and can prioritize and focus their efforts on the areas that receive more negative than positive sentiments.

THIS PAGE INTENTIONALLY LEFT BLANK

V. DISCUSSION

This chapter covers a discussion on the results, approaches taken in the research, recommendations on how this can be used for greater impact, and future work to improve the process and implementation.

A. TAKEAWAY

This thesis research showed the type of insights that can be extracted from free text comments on command climate surveys: underlying topics and sentiments of each comment. NLP techniques were critical and predominantly used for the topic extraction portion, and a pre-trained AI model was used to analyze sentiments (positive, neutral, and negative labels). By combining sentiment analysis

By developing a tool that can extract insights from free text comments and filter the insights through different lenses (question categories only, topics only, and question categories and topics), leaders know which areas need the most attention, can better understand issues by knowing which comments to access for context, and can prioritize and focus their efforts on the areas that receive more negative than positive sentiments.

Ultimately, leaders are empowered to take action on areas that are affecting servicemembers the most, thereby increasing trust between servicemembers and their leadership and also causing a second order effect of improving troop welfare, readiness, and unit cohesion. Additionally, this approach can be replicated to other services and can inspire efforts that will also positively impact talent management and retention.

B. LIMITATIONS AND CHALLENGES

In total, there were three main approaches taken to accomplish the research objectives. The first approach took the entire dataset (after pre-processing) and created 50 clusters (with no additional breakdown prior to clustering). When doing the manual analysis, it became evident for many of the clusters that the top 10 immediate neighbors to each cluster center varied drastically in the topics covered, and thus any topic label applied would be insufficient in representing that cluster.

The second approach sought to improve the results of the first approach by taking the dataset after pre-processing and splitting each comment into sentences before creating 50 clusters. This was done with the intent to represent topics more accurately, since most comments had multiple topics in a single comment. Splitting the dataset down to sentence levels more than doubled the size of the dataset, which was thought to be beneficial as more data is generally good for training models. However, in the manual analysis of the clusters, it became evident that context was lost and like before, topics from immediate neighbors varied drastically, thus any topic label applied would be insufficient in representing that cluster.

The final approach took the pros and cons of the previous two approaches and broke down the dataset based on the questions, which were grouped into categories based on the similar core questions. Each question category dataset varied in size, and thus were clustered into five clusters if the category dataset had less than 300 comments, or ten clusters if the category dataset had more than 300 comments. When working with a single survey, breaking down the dataset by question category will not be necessary, as each question will likely be distinct. This dataset, which was a combination of 11 different surveys from 6 different commands, required grouping questions into similar categories for simplicity and effectiveness.

C. RECOMMENDATIONS

The insights extracted by this tool can provide a competitive advantage in the Marine Corps' ability to understand and address key issues affecting retention and troop welfare at the commander level, as well as higher levels.

The first and most valuable recommendation is to conduct a research project that examines all historical Marine Corps DEOCS data to:

1. Observe trends and changes over time.
2. Overlay with other relevant data, such as operational tempo (deployments, large scale exercises, key inspections such as Field Supply and Maintenance Analysis Office (FSMAO), Commanding General's

Readiness Inspector (CGRI) and safety inspections), personnel turnover, ground flash reports, alcohol related incidents, sexual assault reports (restricted and unrestricted), EO complaints, admittance to behavioral/ mental health care, suicidal ideations, attempts, and actual suicides.

3. Assess differences and similarities amongst units from the ground combat element, aviation combat element, and logistics combat element.
4. Assess differences and similarities from similar combat elements across coasts.

The second recommendation is to make this tool available to commanders at all levels. This can be accomplished by creating a diverse team of people to develop this tool into a program that any commander can have ready access to on their work computers. At a minimum, this team should include people knowledgeable in:

1. Databases
2. Machine learning/deep learning
3. Human computer interface / user experience
4. Marine Corps Enterprise Network limitations
5. The end user

The end user will change depending on the use case that this tool gets applied to. For climate surveys, the end users are commanders who want to get more out of the DEOCS reporting process. For retention data, the end user can include research analysts that are managing the data and reporting to higher levels of leadership on the main factors influencing servicemembers to transition out or stay in the service. For safety survey data, the end user can include the commanders of a unit, safety officers, and higher levels of safety representatives. A new model would need to be built and trained for each new use case to ensure the most accurate results.

The third recommendation comes from an overview of all the comments: there is a heavy emphasis on leadership, the role leadership has, and how accomplishing any mission comes down to the people in the service. The third recommendation is for leaders from all

levels to invest in developing the people in the Marine Corps on more than warfare education, specifically on personal development and people skills, as these directly affect the health, readiness, and longevity of the Marine Corps. This can be achieved by developing a personal development program.

D. FUTURE WORK

With the baseline from this thesis, this tool can easily be applied to any dataset with a large amount of free text comments, such as 1) all Marine Corps DEOCS data (command climate), 2) the Exit Milestone Longitudinal Survey (EMLS) data (retention), 3) Safety Surveys, and 4) any culture/qualitative surveys and studies, to include command climate data from other services. Each of these could be its own research project where deeper analysis can be done.

From the insights gathered with this tool, more questions will be generated to gain better understanding. These insights can serve to iterate improvements and refinements on the questions being asked in a survey much quicker.

Additional worthwhile endeavors to specifically improve the process of creating a prediction model for this tool:

- Create an automated way to 1) read in the PDF documents into Excel files and 2) read in topic label numbers that were determined from manual analysis from Excel documents to the code files.
- Develop a database or other repository to make comments with a certain label accessible, e.g., if a commander wants to look into the personal experiences and perception question category and gain context behind the topic called professionalism/treatment of others, have those comments become readily available via an interface.
- Create a Graphical User Interface (GUI) for the tool, since a GUI will be the biggest difference maker to make this program user friendly for all end users. A different GUI will likely be needed for each use case.

- Add a “mixed” label for sentiment analysis. Similar to Nawab et al. [21], comments with the “mixed” sentiment label can then be further broken to sentence level to get a clearer picture on the sentiment and have a topic predicted for that sentence.
- Explore effects of different vector dimension for Doc2Vec, specifically the influences that a vector dimension for embedding layers have on CNN data input, since the data input needs to be reshaped before the CNN model can use it.
- Optimize parameters for Random Forest and CNN. In this research, default parameters were used for Random Forest, and CNN offers significant flexibility in creating model layers.

E. CHAPTER SUMMARY

This chapter discussed the results, approaches taken in the research, recommendations on how this can be used for greater impact, and future work to improve the process and implementation.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSION

The motivation of the study applied computer science to culture, particularly to assist the Marine Corps with retention and talent management efforts, such as retaining the necessary people for future warfare—people that are innovative, critical thinkers—as well with troop welfare and its connection to military and veteran suicides, and mental health overall. The closest snapshot of the health of Marine Corps culture is collected via command climate surveys, which contain quantitative and free response questions. Senior levels of Marine Corps leadership do not gain value from the free text data because: 1) analysis on free text comments is time consuming and no analysis is done on those comments for the commander; and 2) there is no trend analysis on the aggregate of surveys over time and throughout commands, so they are not commander specific.

The main contribution from this thesis was developing a tool that can help leaders gain value from command climate survey comments. This thesis applied NLP techniques, such as topic extraction and sentiment analysis, on free text comments from command climate surveys to provide a quantitative analysis on the health of culture in Marine Corps units. This approach can be applied to any domain that has large amounts of free text data, as well as replicated to other services, and can inspire efforts that will positively impact talent management and retention.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

- [1] Headquarters, U.S. Marine Corps and D. H. Berger, “Commandant’s planning guidance: 38th Commandant of the Marine Corps,” 2019 [Online]. Available: <https://www.marines.mil/News/Publications/MCPEL/Electronic-Library-Display/Article/1907265/38th-commandants-planning-guidance-cpg/>
- [2] K. Fosher, J. Edwards, R. Lane, E. Tarzi, K. Post, and E. Gauldin, “Marines’ perspectives on various aspects of Marine Corps organizational culture.” Translational Research Group, Center for Advanced Operational Culture Learning, Education Command, 2018 [Online]. Available: [https://www.usmcu.edu/Portals/218/CAOCL/files/MCOOCR Report to PSO 30Mar18_wDem_FINAL.pdf](https://www.usmcu.edu/Portals/218/CAOCL/files/MCOOCR%20Report%20to%20PSO%2030Mar18_wDem_FINAL.pdf)
- [3] Office of People Analytics, “Assessment to solutions,” Defense Equal Opportunity Management Institute [Online]. Available: <https://www.defenseculture.mil/Assessment-to-Solutions/A2S-Home/>.
- [4] Headquarters, U.S. Marine Corps, “Marine Corps Order 5354.1F Marine Corps Prohibited Activities and Conduct (PAC) prevention and response policy,” 2021 [Online]. Available: <https://www.marines.mil/News/Publications/MCPEL/Electronic-Library-Display/Article/1561150/mco-53541f-cancels-mco-53541e-wadmin-ch/>
- [5] S. Arango, “Opinion | The Marine Corps’ culture has to change,” *Task & Purpose*, Apr. 12, 2022 [Online]. Available: <https://taskandpurpose.com/news/marine-corps-retention-talent-management-2030/>.
- [6] J. Berger and G. Packard, “Using natural language processing to understand people and culture,” *American Psychologist*, Dec. 2021, doi: <http://dx.doi.org/10.1037/amp0000882>.
- [7] *CommSafe AI* [Website]. Available: <https://commsafe.ai/>.
- [8] D. Khyani and S. B S, “An interpretation of lemmatization and stemming in natural language processing,” *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, vol. 22, pp. 350–357, Jan. 2021 [Online]. Available: https://www.researchgate.net/publication/348306833_An_Interpretation_of_Lemmatization_and_Stemming_in_Natural_Language_Processing
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003 [Online]. Available: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>

- [10] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Multi-objectives-based text clustering technique using k-mean algorithm," in *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, Jul. 2016, pp. 1–6. doi: 10.1109/CSIT.2016.7549464 [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7549464>
- [11] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." arXiv, May 22, 2014. doi: 10.48550/arXiv.1405.4053 [Online]. Available: <https://arxiv.org/abs/1405.4053>
- [12] H. Dogru, S. Tilki, A. Jamil, and A. Hameed, "Deep learning-based classification of news texts using Doc2Vec model," 2021, p. 96. doi: 10.1109/CAIDA51941.2021.9425290 [Online]. Available: <https://ieeexplore.ieee.org/document/9425290>
- [13] C. M. Cairoli, "Categorization of survey text utilizing natural language processing and demographic filtering." M.S. thesis, Dept. of Operations Research, Monterey, CA; Naval Postgraduate School, 2018 [Online]. Available: <https://calhoun.nps.edu/handle/10945/56109>
- [14] C. Layug, "Extracting major topics from survey text responses using natural language processing," M.S. thesis, Dept. of Operations Research, Monterey, CA; Naval Postgraduate School, 2018 [Online]. Available: <https://calhoun.nps.edu/handle/10945/60425>
- [15] M. P. Salonga, "Natural language processing of short comments from United States Navy survey data," M.S. thesis, Dept. of Operations Research, Monterey, CA; Naval Postgraduate School, 2020 [Online]. Available: <https://calhoun.nps.edu/handle/10945/64875>
- [16] L. Cheng, "Unsupervised topic discovery by anomaly detection," M.S. thesis, Dept. of Operations Research, Monterey, California: Naval Postgraduate School, 2013 [Online]. Available: <https://calhoun.nps.edu/handle/10945/37599>
- [17] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*, New York, NY, USA, Nov. 2009, pp. 375–384. doi: 10.1145/1645953.1646003 [Online]. Available: <https://dl.acm.org/doi/10.1145/1645953.1646003>
- [18] C. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, Art. no. 1, May 2014 [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>

- [19] S. Loria, “Textblob documentation: Release 0.16.0,” Apr. 2020 [Online]. Available: <https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf>
- [20] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, Jun. 2019, pp. 54–59. doi: 10.18653/v1/N19-4010 [Online]. Available: <https://aclanthology.org/N19-4010/>
- [21] K. Nawab, G. Ramsey, and R. Schreiber, “Natural language processing to extract meaningful information from patient experience feedback,” *Appl Clin Inform*, vol. 11, no. 2, pp. 242–252, Mar. 2020, doi: 10.1055/s-0040-1708049 [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7113005/>
- [22] M. S. Neethu and R. Rajasree, “Sentiment analysis in twitter using machine learning techniques,” in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Jul. 2013, pp. 1–5. doi: 10.1109/ICCCNT.2013.6726818 [Online]. Available: <https://ieeexplore.ieee.org/document/6726818>
- [23] K. Yadav, “Cleaning & preprocessing text data by building NLP pipeline,” Medium, Jan. 22, 2022 [Online]. Available: <https://towardsdatascience.com/cleaning-preprocessing-text-data-by-building-nlp-pipeline-853148add68a>.
- [24] S. Pandey and S. K. Pandey, “Applying natural language processing capabilities in computerized textual analysis to measure organizational culture,” *Organizational Research Methods*, vol. 22, no. 3, pp. 765–797, Jul. 2019, doi: 10.1177/1094428117745648.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California