USING SOCIAL MEDIA ACTIVITY TO IDENTIFY PERSONALITY CHARACTERISTICS OF NAVY PERSONNEL

by

Leslie Ward

March 2016

Thesis Co-Advisors:

Man-Tak Shing

Thomas Otani

Approved for public release; distribution is unlimited
This research explores the use of Twitter to determine if the personality characteristics of well-performing Navy personnel can be identified based on their Twitter use. Well-performing Navy personnel are identified by using the publicly-available Navy promotion lists and then those names were used to query Twitter in order to identify possible accounts belonging to these Sailors. Data from those Twitter accounts that could be positively identified as belonging to Navy personnel were then fed into textual analysis software and each user’s level of the personality traits in the Five Factor Model of personality was calculated based on the results previous research. These results and other data were also stored in a graph database in order to make the data easier to query.

Although this research shows that it is possible to successfully calculate a user’s personality based on textual analysis of their Twitter activity, the primary conclusions of this research is that this method is insufficient to identify specific traits that make Navy personnel stand out on Twitter.
USING SOCIAL MEDIA ACTIVITY TO IDENTIFY PERSONALITY CHARACTERISTICS OF NAVY PERSONNEL

Leslie Ward
Lieutenant, United States Navy
B.S., Texas A&M University, 2005

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
March 2016

Approved by: Man-Tak Shing
Thesis Co-Advisor

Thomas Otani
Thesis Co-Advisor

Peter Denning
Chair, Department of Computer Science
ABSTRACT

This research explores the use of Twitter to determine if the personality characteristics of well-performing Navy personnel can be identified based on their Twitter use. Well-performing Navy personnel are identified by using the publicly-available Navy promotion lists and then those names were used to query Twitter in order to identify possible accounts belonging to these Sailors. Data from those Twitter accounts that could be positively identified as belonging to Navy personnel were then fed into textual analysis software and each user’s level of the personality traits in the Five Factor Model of personality was calculated based on the results previous research. These results and other data were also stored in a graph database in order to make the data easier to query.

Although this research shows that it is possible to successfully calculate a user’s personality based on textual analysis of their Twitter activity, the primary conclusions of this research is that this method is insufficient to identify specific traits that make Navy personnel stand out on Twitter.
# Table of Contents

1 Introduction .................................................. 1  
1.1 Motivation ................................................. 1  
1.2 Research Questions ........................................ 2  
1.3 Organization of Thesis ..................................... 2  

2 Background .................................................... 5  
2.1 Personality Traits ......................................... 5  
2.2 Twitter .................................................... 6  
2.3 Graph Databases ........................................... 11  
2.4 Linguistic Inquiry and Word Count ......................... 13  
2.5 Related Work .............................................. 13  

3 Methodology ................................................... 17  
3.1 Identifying Navy Personnel ................................. 17  
3.2 Identifying Twitter Accounts ............................... 17  
3.3 Identifying Personality Characteristics of Each User ... 21  

4 Analysis ......................................................... 29  
4.1 Results .................................................... 29  
4.2 Calculation Anomalies ..................................... 32  

5 Graph Database Storage .................................... 35  
5.1 Graph Database Model .................................... 35  
5.2 Querying the Data ......................................... 39  

6 Conclusion and Future Work ................................. 43  
6.1 Conclusions ............................................... 43  
6.2 Future Work ............................................... 44  

List of References ............................................. 47  

vii
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Tweet</td>
<td>7</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Tweet JSON String</td>
<td>9</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Simple Graph</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Cypher Keywords</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Record Message Traffic Promotion List</td>
<td>18</td>
</tr>
<tr>
<td>Figure 6</td>
<td>All Hands Page Promotion List</td>
<td>19</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Correlations between Personality Traits and LIWC Results</td>
<td>27</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Expected Values of Personality Characteristics</td>
<td>28</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Boxplot of Five Factor Results</td>
<td>30</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Density of Agreeableness Values</td>
<td>31</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Density of Conscientiousness Values</td>
<td>32</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Density of Extroversion Values</td>
<td>33</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Density of Neuroticism Values</td>
<td>33</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Density of Openness to Experience Values</td>
<td>34</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Graph DB Model</td>
<td>36</td>
</tr>
<tr>
<td>Figure 16</td>
<td>User Relationships</td>
<td>38</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Tweet Relationships</td>
<td>39</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Overall Timeline Tree</td>
<td>40</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Timeline Tree and Years</td>
<td>40</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Year and Month Nodes</td>
<td>41</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Month and Day Nodes</td>
<td>41</td>
</tr>
<tr>
<td>Table 1</td>
<td>Example of LIWC Words and Categories</td>
<td>14</td>
</tr>
<tr>
<td>Table 2</td>
<td>List of Search Terms for Identifying Navy Accounts</td>
<td>20</td>
</tr>
<tr>
<td>Table 3</td>
<td>Standard Deviation of Character Traits</td>
<td>30</td>
</tr>
<tr>
<td>Table 4</td>
<td>Correlation between Character Traits and Non-Language Data</td>
<td>34</td>
</tr>
<tr>
<td>Table 5</td>
<td>List of Properties of a User Node</td>
<td>37</td>
</tr>
<tr>
<td>Table 6</td>
<td>List of Properties of a Tweet Node</td>
<td>37</td>
</tr>
</tbody>
</table>
## List of Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BUPERS</td>
<td>Bureau of Personnel</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma Separated Values</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>LIWC</td>
<td>Linguistic Inquiry and Word Count</td>
</tr>
<tr>
<td>NoSQL</td>
<td>Non-Structured Query Language</td>
</tr>
<tr>
<td>RDBMS</td>
<td>Relational Database Management System</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>TAPAS</td>
<td>Tailored Adaptable Personality Assessment System</td>
</tr>
<tr>
<td>USN</td>
<td>United States Navy</td>
</tr>
</tbody>
</table>
Acknowledgments

Special thanks to my thesis advisors, Man-Tak Shing and Thomas Otani. When I came to Dr. Shing with my idea for a thesis, he was enthusiastic in helping me figure out how to make it a viable research idea and he guided me through the entire process. Dr. Otani stepped up at the last minute to be my co-advisor and answered every question that I asked him.

I would also like to thank Michael Atkinson of the NPS Operations Research Department, who walked me through the construction and use of the linear regression model that served as the basis for my calculations.

The initial idea for my thesis came from my work with the Chief of Naval Operations Strategic Studies Group and the idea to use social media profiles for recruitment was introduced and championed by CAPT Reggie Howard.

Thanks to LT Patrick Gillen, whose work with Twitter to identify OPSEC violations first inspired me to look at social media and who provided me with invaluable guidance for identifying Navy Twitter accounts and working with the Twitter API.

Thanks to LT Brian Crawford, who served as my sounding board to discuss problems and to listen to my rants when things weren't going well.

Last, but certainly not least, thanks to my family. My mother first introduced me to computer science and showed me how much fun programming is, which changed the course of my life. Both of my parents were always enthusiastic and interested to hear about what I was doing in my thesis and what I was learning in my classes. I appreciate my brothers and sister for not only giving me nieces and nephews, but also for being willing to drop everything and come see me whenever I make it back to Texas for a visit.
CHAPTER 1:  
Introduction  

Nearly 65% of Americans use social media, according to the 2015 survey by Pew Research Center; in the 18–29 age range typically targeted for military recruitment, the number jumps to 90% [1]. The list of social media platforms is constantly growing. The most commonly used platforms are YouTube, Facebook, Google+, and Twitter; others include LinkedIn, Tumblr, Instagram, and Pinterest [2]. Social media is the third most common entertainment choice for Americans aged 16–24, behind television and hanging out [2]. People use social media to interact with friends, family members and celebrities. They post about the big events and little events in their lives. They provide their opinions on politics, world news, movies and TV, and sports. Americans spend an average of nearly two hours a day on social media; for 16–35 year-olds, that number is even higher [3].

Most social media platforms have the right, as laid out in their Terms of Service, to provide users’ data to third-party sources, typically marketing firms. These third-party companies use data mining tools to identify targets for advertising and to determine trends, because there is so much useful information in a user’s data. For example, LinkedIn has marketed its platform as a tool for both employers to find candidates with specific skills and for job-seekers to find employment.

1.1 Motivation
The U.S. Navy has a recruitment goal of 37,000 new active duty members in 2016 [4]. The current Navy recruiting process has multiple ways to identify potential new recruits; the process is known as prospecting. Recruiters visit schools, malls, parks, sporting events and unemployment offices to seek new prospects; recruiters attend 32,000 high schools and 5,000 colleges every year to find those recruits [4]. They canvas schools and current applicants for referrals. The names and contact information of the prospects are then used for follow-up contact. The Navy Recruiting Manual recommends the telephone as the best way to make the initial contact [5]. The manual also recommends mail-outs and social media networks as alternate ways to contact potential recruits. The goal of these contacts is to set up an appointment for an interview between the recruiter and the prospect.
Despite the widespread use of social media by other companies for targeted marketing and job placement, the U.S. Navy has not embraced its use beyond basic non-targeted marketing. The Navy Recruiting Manual only has one paragraph on using social media for recruiting purposes, and it focuses on how to document the contact; it does not discuss how to discover prospects [5]. This is further evidence that the U.S. military is focusing on social media as an additional advertising tool instead of as a recruiting tool.

Navy Recruiting Command lists its recruiting priorities as follows: Medical officers, Chaplains, SEALs, Navy Special Warfare, Navy Special Operations, Special Warfare Combatant-Craft Crewmen, Explosive Ordnance Disposal, Diver, Hospital Corpsmen, and Reserves [4]. All of these jobs require some kind of special qualifications or aptitudes. However, the methods that recruiters have now are insufficient to identify potential prospects with the right qualifications or aptitudes and the personality characteristics necessary to be a successful Sailor.

1.2 Research Questions
This research explores the use of social media, specifically Twitter, to determine if the personality of well-performing Navy personnel can be identified based on their Twitter use and if so, what other useful information can be determined that might differentiate a well-performing Navy Twitter user from a non-Navy Twitter user. The term "well-performing" is used to indicate those Sailors whose contribution to the Navy is positive; this research uses selection for promotion as a proxy for "well-performing."

By answering these questions, this research takes the first step in determining whether a tool to identify future recruits based on their Twitter activity would be both feasible and useful. This notional tool would allow recruiters to identify potential prospects with the right aptitude who would not otherwise consider a career in the Navy, and target them for recruitment.

1.3 Organization of Thesis
Chapter 2 provides background information on the study and characterization of personality traits, the Twitter social media platform, graph databases, the Linguistic Inquiry and Word Count (LIWC) software, and related research in this area. Chapter 3 covers the methodology
used to identify the accounts of Navy personnel and the equations used to identify each user’s personality characteristics. Chapter 4 contains the findings of the research. Chapter 5 explains the model used to store the data and identifies some of the questions that can be answered by querying the data. Chapter 6 contains the conclusions and recommendations for future work on this topic.
CHAPTER 2:
Background

This chapter provides background information on the different topics addressed in this thesis, including the Five Factor Model of personality, the Twitter social media platform, graph databases, and the Linguistic Inquiry and Word Count (LIWC) software.

2.1 Personality Traits
The field of psychology has been attempting to quantify humans via personality for at least the last century [6]. Many models have been proposed over the years, but few have withstood additional testing. However, the Five Factor Model of personality traits, also known as the Big Five, has been shown to be robust against different methods of testing and is the most commonly used approach for personality identification in psychology today. The personality traits identified in this thesis are based on the Five Factor Model.

2.1.1 The Five Factor Model
The central idea of the Five Factor Model is that all personality traits can be categorized into one of the five factors, and any person can be described by their rating for each of the factors [7]. The five factors are Agreeableness, Conscientiousness, Extroversion, Neuroticism, and Openness to Experience [8].

One weakness in the Five Factor Model is that there is no official definition of the terms; however, similar words are used to describe each of the factors across much of the research [9].

The five factors are:

- **Agreeableness**, described with terms such as trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness
- **Conscientiousness**, described with terms such as competence, order, dutifulness, achievement striving, self-discipline, and deliberation
• **Extroversion**, described with terms such as warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotions
• **Neuroticism**, described with terms such as anxiety, anger, depression, self-consciousness, impulsiveness, and vulnerability
• **Openness to Experience**, described with terms such as fantasy, aesthetics, feelings, actions, ideas, and culture [9]

There is no standard scale used to describe these factors; this research uses the same 0–1 scale as seen in [10].

### 2.2 Twitter

Twitter is a social media platform designed for microblogging; all posts are limited to 140 characters. Twitter provides a medium for users to post about their lives, activities, and opinions. Users are referenced by both a unique *screen name* chosen by the user and a unique user identification number assigned by Twitter. Although a user can change their screen name, their user identification number remains the same. Screen names are displayed on the pages through the Twitter site, and their identification numbers are available in the HTML code for a page. Users have the option to set their accounts to *protected*, which limits public access to any of their activity beyond basic profile data; without this restriction, all posts are available to the public.

Twitter posts are known as *tweets* or *statuses* and are also assigned unique identification numbers. People who are subscribed to a user’s posts are known as *followers*. Users can *favorite* or *retweet* a post to indicate their support of that tweet. Within a tweet, a user can use a word or phrase (without spaces), called a *hashtag* and identified by the character #, which links that post to any other tweet containing the same hashtag. Twitter displays the most commonly used hashtags on its main page to show what is trending at any time. Users can embed photos or videos in their tweets. Other users can be referenced in a tweet by using the character @ and a screen name; these references are either a *reply*, where the tweet is a direct response to another tweet, or a *mention*. User mentions are more commonly used by users who are trying to get the attention of a celebrity. Although anyone can create an account using any name, celebrity accounts are *verified* by Twitter as actually belonging to the celebrity they are claiming to represent. An example of a tweet can be seen in Figure 1.
2.2.1 Twitter API

Twitter is accessible for developers using an application programming interface (API). The Twitter API is divided into three categories: the REST API, the Streaming API, and the Streaming Firehose [11]. The REST API provides access to the Twitter data stream for individual transactions such as posting a tweet, reading a user profile or identifying followers. The Streaming API and the Streaming Firehose are both used for persistent connection transactions such as reading tweets over a period of time; the difference is in the
amount of the overall Twitter traffic that can be accessed [11]. This work exclusively used the REST API.

The API provides data in four different object types—*Tweets, Users, Entities, and Places*—using JavaScript Object Notation (JSON) strings. An example of a tweet in a formatted JSON string is shown in Figure 2. The types and formatting of information provided by Twitter in the JSON string is not the same for every object, even within the same object type; generally, if a field is empty or null, it is not returned as part of the JSON string at all. Twitter programmers also change the included metadata and formatting as they see fit and warn that developers’ applications need to be able to tolerate the changes [11].

**Tweet Object**
A Tweet object provides both the text of the tweet and the metadata about the tweet. The fields that may be included in a Tweet object as of the time of data collection for this research are:

- **contributors**: A collection of users who contributed to the authorship of the tweet.
- **coordinates**: The latitude and longitude of the tweet.
- **created_at**: The date and time when the tweet was created.
- **favorite_count**: The number of users who have favorited this tweet.
- **id**: A unique integer identifier for the tweet.
- **in_reply_to_screen_name**: If the tweet is a reply to another tweet, this contains the screen name of the original author.
- **in_reply_to_status_id**: If the tweet is a reply to another tweet, this contains the ID number of the original tweet.
- **in_reply_to_user_id**: If the tweet is a reply to another tweet, this contains the ID number of the original author.
- **lang**: The language of the tweet text, if it can be determined.
- **place**: A Place object as described in Section 2.2.1.
- **retweeted_status**: If the tweet is a retweet, this field contains a Tweet object representing the original tweet.
- **source**: The application used to post the tweet.
- **text**: The actual text of the tweet.
- **user**: A User object, as described in Section 2.2.1, representing the user who posted
Figure 2: Example of the Tweet from Figure 1 as a formatted JSON string.

this tweet.

• **user_mentions**: A list of the users referenced in the Tweet, with shortened User objects for each user.

**User Object**

A User object provides the metadata about the user. The fields that may be included in a User object as of the time of data collection for this research are:

• **created_at**: The date and time that the user account was created.
• *description*: The user’s free-text description of their account.
• *entities*: One or more Entity objects as described in Section 2.2.1.
• *favorites_count*: The number of tweets the user has favorited.
• *followers_count*: The number of followers the user has.
• *friends_count*: The number of accounts this user is following.
• *geo_enabled*: Indicates if the user has allowed geo-tagging of their tweets.
• *id*: A unique integer identifier of the user.
• *lang*: The default language for the user’s interface.
• *location*: The user-defined location in a string format.
• *name*: The name of the user.
• *protected*: A Boolean variable that indicates if the user has protected their account. For a protected account, only the information in the User object JSON string is available; all other information, including tweets and followers, is only available to those that the user has explicitly granted permission to.
• *screen_name*: The screen name of the user.
• *status*: A Tweet object containing the user’s most recent status.
• *statuses_count*: The number of tweets, including replies and retweets, that the user has posted.
• *url*: A URL provided by the user.

**Entity Object**
An Entity object provides additional metadata about a tweet or user. The fields that may be included in an Entity object as of the time of data collection for this research are:

• *hashtags*: A list of the hashtags contained in the object.
• *media*: A representation of the media elements in the object.
• *url*: A list of the URLs included in the object.
• *user_mentions*: A list of the users referenced in the Tweet, with shortened User objects for each user.

**Place Object**
A Place object provides additional metadata about a place. The place can be either the location where the tweeted was posted from or a place mentioned in the tweet. The fields
that may be included in a Place object as of the time of data collection for this research are:

- **bounding_box**: A set of coordinates that describe the bounds of the place.
- **country**: The country name of the place.
- **country_code**: A shortened form of the country name.
- **full_name**: The full name of the place in human-readable form.
- **id**: A unique string representing the place.
- **name**: A shortened form of the human-readable name.
- **place_type**: The type of place.

### 2.2.2 Access and Limitations

There are wrappers available for the Twitter API in many different programming languages in order to make it easier for developers to use the API. This work used Python and the wrapper `python-twitter` to access the API and for further data processing.

Access to the Twitter API requires a Twitter account and registration for a Twitter App Token, both of which are free and only require an email address in order to register. Access to the REST and Streaming APIs are also free; however, both have limitations on their use. The REST API is limited to 180 queries in a 15-minute window; this was a hindrance to data collection for this thesis as it greatly increased the time required to gather the necessary information. The Streaming API provides real-time access to tweets, but only a fraction of the total at any point—generally 1%, though it can be higher during low-traffic periods [12]. The only way to get access to 100% of tweets in real time is via the Twitter Firehose, which is a paid service.

### 2.3 Graph Databases

Relational database management systems (RDBMS) are the most common way that data is stored in a database. Data in an RDBMS is stored in relational tables and accessed via a Structured Query Language (SQL) [13]. Database management systems that do not use relational tables or SQL are collectively referred to as NoSQL databases. A graph database management system is one of several types of NoSQL databases, in which data is stored and queried using a graph model and graph theory, as opposed to the tables and cross-product queries of an RDBMS [13].
A graph is a set of vertices or nodes that are connected by edges. The edges may or may not be directional. Graph databases prioritize the relationships between data and allow complicated queries that follow through multiple connections, which are memory and processing-intensive in relational databases. Almost any data that can be modeled using an RDBMS can also be modeled in a graph database, but graph databases are especially useful for storing data such as business or social networks [13].

This work used the graph database program Neo4j and the query language Cypher to create and query the database. Neo4j uses nodes, relationships, properties, and labels as its basic building blocks. Nodes in Neo4j are equivalent to nodes or vertices in graphs. Relationships are equivalent to edges in graphs and are used to connect nodes. Both relationships and nodes can have properties, which add more detail to them. Labels are used to group nodes or relationships by type [13]. An example of a simple Neo4j graph is shown in Figure 3.

![Figure 3: Simple graph pattern, where the blue circles are nodes with the label Person and the property name. The nodes are connected to each other with the relationship KNOWS. Source: I. Robinson, J. Webber, and E. Eifrem, Graph Databases: New Opportunities for Connected Data, 2nd ed. Sebastopol, CA: O’Reilly Media, Inc, 2015.](image)

Cypher is similar in format to SQL, the language used to query relational databases, but uses different reserve words. Figure 4 shows the key words available in Cypher. A simple question for the graph in Figure 3 would be to find out who the Person named Jim knows. The Cypher query for that question is:

```cypher
MATCH (a:Person)-[:KNOWS]->(b:Person)
WHERE a.name = 'Jim'
RETURN b
```
which should return two nodes, a Person named Ian and a Person named Emil. Cypher can also be used to answer much more complicated questions.

<table>
<thead>
<tr>
<th>MATCH</th>
<th>Identifies data matching the specified pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>RETURN</td>
<td>Returns the data to the client</td>
</tr>
<tr>
<td>WHERE</td>
<td>Provides criteria for filtering pattern matching results.</td>
</tr>
<tr>
<td>CREATE</td>
<td>Creates nodes and relationships.</td>
</tr>
<tr>
<td>CREATE UNIQUE</td>
<td>Creates nodes and relationships.</td>
</tr>
<tr>
<td>MERGE</td>
<td>Ensures that the supplied pattern exists in the graph, either by reusing existing nodes and relationships that match the supplied predicates, or by creating new nodes and relationships.</td>
</tr>
<tr>
<td>DELETE</td>
<td>Removes nodes, relationships, and properties.</td>
</tr>
<tr>
<td>SET</td>
<td>Sets property values.</td>
</tr>
<tr>
<td>FOREACH</td>
<td>Performs an updating action for each element in a list.</td>
</tr>
<tr>
<td>UNION</td>
<td>Merges results from two or more queries.</td>
</tr>
<tr>
<td>WITH</td>
<td>Chains subsequent query parts and forwards results from one to the next. Similar to piping commands in Unix.</td>
</tr>
<tr>
<td>START</td>
<td>Specifies one or more explicit starting points—nodes or relationships—in the graph.</td>
</tr>
</tbody>
</table>

Figure 4: Cypher Keywords and Descriptions. Adapted from: I. Robinson, J. Webber, and E. Eifrem, *Graph Databases: New Opportunities for Connected Data*, 2nd ed. Sebastopol, CA: O'Reilly Media, Inc, 2015.

### 2.4 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a software tool used to analyze text. Given a sample of text, LIWC counts the occurrences of different types of words as defined by a pre-loaded or user-defined dictionary of words and categorization of those words [14]. Table 1 shows a list of categories and example words that fall within those categories. The results for each category are returned as a percentage of the overall number of words in the sample. Words can fall into multiple categories or not be included in any category, so the sum of the percentages for the categories will not equal 100%. This work used LIWC2015 with the pre-loaded dictionary; no user-defined dictionaries were used.

### 2.5 Related Work

Many studies have been done to correlate personality and job performance. Although research prior to 1990 generally was unable to determine any correlation, more reliable
correlations have been determined since the general acceptance of the Five Factor Model and its use in these studies [6]. Conscientiousness has been consistently shown as the most important factor in overall job performance. The other factors’ importance in job performance is based on the type of job [6].

In [15], researchers demonstrated that military personnel who showed low levels of depression and homesickness and who adjusted to the military lifestyle more easily also showed low levels of Neuroticism and higher levels of Extroversion and Openness to Experience. They also showed that those who were rated as effective by both their direct superior and by themselves showed higher levels of Conscientiousness than those not rated as effective. These studies show that identifying personality traits according to the Five Factor Model can provide useful information for identifying possible recruits.

In [16], researchers examined multiple models to automatically identify personality based on written input; that research was extended to include both essays and recorded snippets of conversations in [17]. In [18], researchers were able to predict user’s personality in the Five Factor Model based on their Facebook activity. That work was extended to Twitter in [10]. In [19], users were classified by both personality and profession based on their Twitter activity. These papers show that it is possible to determine a person’s personality traits based on their writing and social media activity. This research uses similar methodology, focusing specifically on Navy personnel, in order to determine if more useful information

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal pronouns</td>
<td>I, them, her</td>
</tr>
<tr>
<td>Impersonal pronouns</td>
<td>it, it’s, those</td>
</tr>
<tr>
<td>Articles</td>
<td>a, an, the</td>
</tr>
<tr>
<td>Prepositions</td>
<td>to, with, above</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>am, will, have</td>
</tr>
<tr>
<td>Common Adverbs</td>
<td>very, really</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>and, but, whereas</td>
</tr>
<tr>
<td>Negations</td>
<td>no, not, never</td>
</tr>
</tbody>
</table>

Table 1: Example of LIWC words and categories.
can be determined based on their Twitter activity and personality.

The use of Navy promotion lists to identify Twitter accounts was previously done in [20]; this research uses the same methodology to identify accounts for further processing.
CHAPTER 3:  
Methodology

This chapter explains the methodology used to identify the accounts of Navy personnel, and the equations used to identify each user’s personality characteristics.

3.1 Identifying Navy Personnel

The data collection phase of this research began with identifying well-performing Navy personnel, defined as those who have been selected for a promotion to higher rank. All Navy promotion lists are published online and are publicly available; there are multiple different formats and sites with the data. Officer promotion lists are disseminated via record message traffic to all Navy units and posted to the Navy Bureau of Personnel (BUPERS) website in text format. Figure 5 shows the beginning of an officer promotion message. Enlisted promotion lists are generally posted in PDF format to the Navy All Hands page at www.navy.mil 24 hours after commands have been notified. Figure 6 shows an example of an enlisted promotion list.

Promotion lists are released twice a year for the pay grades E-4 through E-6 and once a year for pay grades E-7 through E-9 and O-3 through O-6. E-1 through E-3 and O-1 through O-2 promotions are based solely on time-in-grade and no lists of those promoted are published. O-7 and above promotions are based on assignments to a specific job and are announced as necessary throughout the year. This work used all of the promotion lists from Fiscal Year 2015, between October 2014 and September 2015, for the pay grades E-4 through E-8 and O-3 through O-5. There were a total of 54,580 names on all of these lists combined.

3.2 Identifying Twitter Accounts

After compiling the list of names, I used a python script and the python-twitter wrapper for the Twitter API to search for each of the names on Twitter. Each search request returned up to 100 user profile strings in JSON format, which were then converted to a comma-separated string and stored in a comma-separated values (CSV) file. Because of the large number of names that were searched for and the Twitter REST API query rate limits,
this process took approximately 80 hours to complete and returned approximately 280,000 Twitter accounts.

Due to the nature of this research, it was important that only accounts actually belonging to Navy personnel were included in the data collection and analysis. Because of the large number of accounts, it was not feasible to look at each account individually to verify whether or not it actually belonged to a member of the Navy. Each user profile was instead run through a script that checked the JSON string for matches from a list of key words, including references to the Navy, Navy titles, and common Navy locations; the full list of key words is shown in Table 2. These terms were case-insensitive in the search. This returned 6,884
Each of the remaining user profile CSV strings was prepended with an m—to signify maybe—and then examined manually in an Excel spreadsheet. Accounts that belonged to users who were obviously in the Navy were marked with a y and accounts that belonged to users who were obviously not in the Navy were marked with an n. Obvious disqualifiers included: having a foreign location that was not one of the known overseas military locations; profile name not matching the requested search name; and profile descriptions mentioning an occupation that was not the U.S. Navy. Protected accounts were also excluded, whether or not the user could be identified as being in the Navy, because the protected status prevents access to their tweets, which is what this research was looking for.

This step of the process identified 380 accounts that obviously belonged to Navy personnel, 5,839 accounts that obviously did not belong to Navy personnel, and 665 accounts that could not be categorized either way. These 665 accounts were then examined manually.
<table>
<thead>
<tr>
<th>Navy</th>
<th>USN</th>
<th>naval</th>
</tr>
</thead>
<tbody>
<tr>
<td>sea</td>
<td>sub</td>
<td>pilot</td>
</tr>
<tr>
<td>aviat</td>
<td>Sailor</td>
<td>USS</td>
</tr>
<tr>
<td>petty</td>
<td>chief</td>
<td>SWO</td>
</tr>
<tr>
<td>military</td>
<td>Newport</td>
<td>Groton</td>
</tr>
<tr>
<td>Washington</td>
<td>Annapolis</td>
<td>Norfolk</td>
</tr>
<tr>
<td>Virginia Beach</td>
<td>Va Beach</td>
<td>Charleston</td>
</tr>
<tr>
<td>King’s</td>
<td>Jacksonville</td>
<td>Mayport</td>
</tr>
<tr>
<td>Pensacola</td>
<td>Millington</td>
<td>Corpus</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>San Diego</td>
<td>Monterey</td>
</tr>
<tr>
<td>Everett</td>
<td>Bremerton</td>
<td>Bangor</td>
</tr>
<tr>
<td>Pearl H</td>
<td>Yoko</td>
<td>Sasebo</td>
</tr>
<tr>
<td>Rota</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: List of search terms used to identify Navy Twitter accounts

by opening their Twitter page and searching their pictures, tweets, and who they were following to determine if they were actually Navy personnel. As seen in [20], many of the Navy accounts could be easily identified by profile pictures of the user in uniform or tweets about Navy activities. Examining who a user was following generally did not provide any useful data; following one or more of the official Navy accounts was not a enough in itself to declare the account as belonging to a Navy member, though it was combined with other individually inconclusive factors. When there was any doubt about whether the user was in the Navy, I erred on the side of caution and excluded them. The final number of verified Navy personnel user accounts was 500.

3.2.1 Collecting Tweets

For each of the 500 verified Navy accounts, I queried the Twitter REST API for the most recent 2000 tweets, including retweets of others’ tweets; for those users with fewer than 2000 tweets, their full tweet history was returned. The earliest tweet came from 7 June 2008, and the longest time between a user’s most recent tweet and their first or 2000th tweet—whichever was later—was seven years and two months. There were a total of 72,678 tweets, with an average of 145 tweets per user and a median of 184 tweets per user.
3.3 Identifying Personality Characteristics of Each User

Once the data was collected and stored in the database, it was analyzed using LIWC2015 as described in Section 2.4. Each user’s tweets were analyzed together as an overall corpus.

To determine a user’s level for each of the five personality factors, I used the basic linear regression equation

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... \beta_m x_{im} + \epsilon_i, \]  

(3.1)

where \( Y_i \) represents the \( i \)th user’s level of a certain character trait \( Y \), \( x_{ij} \) is the value of the \( j \)th independent variable for user \( i \) as determined by LIWC, and \( \beta_j \) is the coefficient of the \( j \)th independent variable, as calculated using Equation 3.3.

Each of the five character traits uses a different set of independent variables, based on the work by Golbeck et al. in [10], which determined the correlation coefficient between a user’s level of a certain trait and the results of using LIWC on their Twitter corpus. These correlation coefficient are shown in Figure 7.

For Extroversion, the LIWC categories that showed significant correlation were: Social Processes, Family, Health, Question Marks and Parentheses. For Agreeableness, the LIWC categories that showed significant correlation were: You, Causation, Ingestion, Achievement, and Money. For Conscientiousness, the LIWC categories that showed significant correlation were: You, Auxiliary Verbs, Future Tense, Negations, Negative Emotions, Sadness, Cognitive Mechanisms, Discrepancy, Feeling, Work, Death, Fillers, Commas, Colons and Exclamation Marks. For Neuroticism, the LIWC categories that showed significant correlation were: Hearing, Feeling, Religion and Exclamation Marks. For Openness to Experience, the LIWC categories that showed significant correlation were: Articles, Quantifiers, Causation, Certainty, Biological Processes, Body, Work, Exclamation Marks, and Parentheses.

For each character trait, a matrix was constructed in which each row represented a single user, the first column consisted of 1’s to represent the lack of \( x \) value for \( \beta_0 \) and each subsequent column represented one of the significant LIWC categories for that character trait. For example, if User 1 has a score of 1.37 for You, 0.8 for Causation, 0.31 for Ingestion, 1.68 for Achievement and 0.44 for Money and User 2 has a score of 6.27 for You, 0.85 for
Causation, 0.78 for Ingestion, 1.56 for Achievement and 0.17 for Money, then the first lines of the matrix for Agreeableness would be:

\[
\begin{bmatrix}
1 & 1.37 & 0.8 & 0.31 & 1.68 & 0.44 \\
1 & 6.27 & 0.85 & 0.78 & 1.56 & 0.17 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{bmatrix}
\]

The matrix for Agreeableness will hereafter be referred to as \( \hat{X}_A \).

The vector of \( \beta \) values for Agreeableness can be written as \( \hat{\beta}_A \) and the vector of \( Y \) values for Agreeableness can be written as \( \hat{Y}_A \), leading to the equation

\[
\hat{Y} = \hat{\beta}_A \hat{X}_A + \epsilon.
\] (3.2)

\( \hat{X}_A \) consists of known values as computed by LIWC. \( \hat{Y}_A \) represents the values I am trying to calculate. \( \hat{\beta}_A \) can be calculated using the formula for ordinary least squares estimation,

\[
\hat{\beta}_A = (\hat{X}_A^T \hat{X}_A)^{-1} \hat{X}_A^T \hat{Y}_A,
\] (3.3)

where \( T \) indicates the transpose matrix and \(-1\) indicates the inverse matrix. \((\hat{X}_A^T \hat{X}_A)^{-1}\) can be calculated from the known values, but \( \hat{Y}_A \) is unknown and therefore must be estimated from the expected value and standard deviation of Agreeableness, hereafter referred to as \( Y_A \), as well as the expected value, standard deviation, and correlation for each \( x_j \). The expected value and standard deviation of \( Y_A \) and the correlation between \( Y_A \) and \( x_j \) are taken from [10], as seen in Figures 7 and 8.

Because that work did not include the expected value or the standard deviation for each \( x_j \), those values are computed using the data in this research. This is possible due to the assumption that the two data sets represent a sufficiently similar population.

The second half of Equation 3.3, \( \hat{X}_A^T \hat{Y}_A \), can be written as
Using the Pearson product-moment correlation coefficient,

\[ \sum_{i=1}^{n} x_{ij} Y_{Ai} = \rho(y, x_j)(n - 1)SD_y SD_{x_j} + n\bar{Y}_A \bar{x}_j; \]

therefore

\[ \hat{X}_A^T \hat{Y}_A = \begin{bmatrix} n\bar{Y}_A \\ \sum_{i=1}^{n} x_{i1} Y_{Ai} \\ \sum_{i=1}^{n} x_{i2} Y_{Ai} \\ \vdots \\ \sum_{i=1}^{n} x_{im} Y_{Ai} \end{bmatrix}. \]

Using \( n=500 \) — the number of user accounts in the data set — and substituting the known values for Agreeableness gives

\[ \hat{X}_A^T \hat{Y}_A = \begin{bmatrix} (500)(0.697) \\ (0.364)(499)(0.162)(1.855) + (500)(0.697)(2.295) \\ (-0.258)(499)(0.162)(1.298) + (500)(0.697)(1.1148) \\ (0.247)(499)(0.162)(0.937) + (500)(0.697)(0.64174) \\ (-0.240)(499)(0.162)(1.172) + (500)(0.697)(1.377) \\ (-0.259)(499)(0.162)(0.782) + (500)(0.697)(0.4996) \end{bmatrix} = \begin{bmatrix} 348.5 \\ 854.25 \\ 360.27 \\ 242.35 \\ 457.18 \\ 157.75 \end{bmatrix}. \]
Calculating out Equation 3.3 gives

\[
\hat{\beta}_A = \begin{bmatrix}
0.702 \\
0.0262 \\
-0.0260 \\
0.0331 \\
-0.0250 \\
-0.0455
\end{bmatrix},
\]

and Equation 3.1 for Agreeableness can be rewritten as

\[
Y_{Ai} = 0.702 + (0.262)x_{iYou} + (-0.0260)x_{iCausation} + (0.0331)x_{iIngestion} \\
+ (-0.0250)x_{iAchievement} + (-0.0455)x_{iMoney} + \epsilon_i,
\]

which can then be applied to each user, resulting in an Agreeableness value for that user. Using the same equations and steps for the other four factors produces the equations:

\[
Y_{Ni} = 0.224 + (0.0720)x_{iHearing} + (0.0908)x_{iFeeling} \\
+ (0.157)x_{iReligion} + (0.0192)x_{iExclamationMarks} + \epsilon_i,
\]

\[
Y_{Ci} = 0.634 + (0.0378)x_{iYou} + (-0.00136)x_{iAuxVerbs} + (-0.0323)x_{iFuture} \\
+ (-0.0308)x_{iNegations} + (0.0225)x_{iNegEmotions} + (-0.0961)x_{iSadness} \\
+ (0.0104)x_{iCogMechanisms} + (-0.0909)x_{iDiscrepancy} + (-0.0241)x_{iFeeling} \\
+ (0.0242)x_{iWork} + (-0.187)x_{iDeath} + (-0.268)x_{iFillers} \\
+ (-0.227)x_{iCommas} + (0.104)x_{iColons} + (0.0148)x_{iExclamationMarks} + \epsilon_i,
\]
\[
Y_{Oi} = 0.583 + (0.0208)x_{i,\text{Articles}} + (0.0153)x_{i,\text{Quantifiers}} + (0.0294)x_{i,\text{Causion}}
\]
\[
+ (0.0373)x_{i,\text{Certainty}} + (0.0110)x_{i,\text{BioProcesses}} + (-0.0150)x_{i,\text{Body}}
\]
\[
+ (0.0249)x_{i,\text{Work}} + (-0.00865)x_{i,\text{ExclamationMarks}} + (-0.0436)x_{i,\text{Parentheses}} + \epsilon_i,
\]

and

\[
Y_{Ei} = 0.481 + (0.00999)x_{i,\text{SocialProcesses}} + (0.141)x_{i,\text{Family}} + (-0.0861)x_{i,\text{Health}}
\]
\[
+ (0.0321)x_{i,\text{QuestionMarks}} + (-0.0531)x_{i,\text{ParenthesesMarks}} + \epsilon_i.
\]

By applying these five equations to the LIWC results, the level of each personality factor can be calculated for each user.

### 3.3.1 Other Statistical Analyses

For each user, statistics were also collected for items not measured by LIWC. These non-language data points are:

- **Followers**: the number of other accounts that are following a user.
- **Following**: the number of other accounts that a user is following.
- **Favorites**: the number of tweets that the user has marked as a favorite.
- **Tweets**: the number of tweets that a user has posted that are included in this data set as described in Section 3.2.
- **Retweets**: the number of a user’s tweets that were a retweet of another user’s post.
- **Replies**: the number of a user’s tweets that were a reply to another user’s post.
- **Hashtags**: the total number of hashtags that a user has posted.
- **Media**: the total number of photos and videos that a user has posted.
- **Words per tweet**: the total number of words of a user normalized by the number of tweets the user has posted.
- **Retweets per tweet**: the number of a user’s retweets normalized by the number of tweets the user has posted.
- **Replies per tweet**: the number of a user’s replies normalized by the number of tweets
the user has posted.

- **Hashtags per tweet**: the number of hashtags a user has included normalized by the number of tweets the user has posted.

- **Media per tweet**: the number of photos and videos that a user has included normalized by the number of tweets the user has posted.

- **Followers/Following**: the ratio between the number of followers a user has and the number of accounts a user is following.
Figure 7: Pearson correlation values between feature scores and personality scores. Significant correlations are shown in bold for $p < 0.05$. Only features that correlate significantly with at least one personality trait are shown.


<table>
<thead>
<tr>
<th>Language Feature</th>
<th>Examples</th>
<th>Extraverted</th>
<th>Agreeable</th>
<th>Conscientious</th>
<th>Neurotic</th>
<th>Open-minded</th>
</tr>
</thead>
<tbody>
<tr>
<td>“You”</td>
<td>(you, your, thou)</td>
<td>0.068</td>
<td>0.364</td>
<td>0.252</td>
<td>-0.212</td>
<td>-0.020</td>
</tr>
<tr>
<td>Articles</td>
<td>(a, an, the)</td>
<td>-0.039</td>
<td>-0.139</td>
<td>-0.071</td>
<td>-0.154</td>
<td>0.396</td>
</tr>
<tr>
<td>Auxiliary Verbs</td>
<td>(am, will, have)</td>
<td>0.033</td>
<td>0.042</td>
<td>-0.284</td>
<td>0.017</td>
<td>0.045</td>
</tr>
<tr>
<td>Future Tense</td>
<td>(will, gonna)</td>
<td>0.227</td>
<td>-0.100</td>
<td>-0.286</td>
<td>0.118</td>
<td>0.142</td>
</tr>
<tr>
<td>Negations</td>
<td>(no, not, never)</td>
<td>-0.020</td>
<td>0.048</td>
<td>-0.374</td>
<td>0.081</td>
<td>0.040</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>(few, many, much)</td>
<td>-0.002</td>
<td>-0.057</td>
<td>-0.089</td>
<td>-0.051</td>
<td>0.238</td>
</tr>
<tr>
<td>Social Processes</td>
<td>(mate, talk, they, child)</td>
<td>0.262</td>
<td>0.156</td>
<td>0.168</td>
<td>-0.141</td>
<td>0.084</td>
</tr>
<tr>
<td>Family</td>
<td>(daughter, husband, aunt)</td>
<td>0.338</td>
<td>0.020</td>
<td>-0.126</td>
<td>0.096</td>
<td>0.215</td>
</tr>
<tr>
<td>Humans</td>
<td>(adult, baby, boy)</td>
<td>0.204</td>
<td>-0.011</td>
<td>0.055</td>
<td>-0.113</td>
<td>0.251</td>
</tr>
<tr>
<td>Negative Emotions</td>
<td>(hurt, ugly, nasty)</td>
<td>0.054</td>
<td>-0.111</td>
<td>-0.268</td>
<td>0.120</td>
<td>0.010</td>
</tr>
<tr>
<td>Sadness</td>
<td>(crying, grief, sad)</td>
<td>0.154</td>
<td>-0.203</td>
<td>-0.253</td>
<td>0.230</td>
<td>-0.111</td>
</tr>
<tr>
<td>Cognitive Mechanisms</td>
<td>(cause, know, ought)</td>
<td>-0.008</td>
<td>-0.089</td>
<td>-0.244</td>
<td>0.025</td>
<td>0.140</td>
</tr>
<tr>
<td>Causation</td>
<td>(because, effect, hence)</td>
<td>0.224</td>
<td>-0.258</td>
<td>-0.155</td>
<td>0.004</td>
<td>0.264</td>
</tr>
<tr>
<td>Discrepancy</td>
<td>(should, would, could)</td>
<td>0.227</td>
<td>-0.055</td>
<td>-0.292</td>
<td>0.187</td>
<td>0.103</td>
</tr>
<tr>
<td>Certainty</td>
<td>(always, never)</td>
<td>0.112</td>
<td>-0.117</td>
<td>-0.069</td>
<td>-0.074</td>
<td>0.347</td>
</tr>
<tr>
<td>Perceptual Processes</td>
<td>(listen, hearing)</td>
<td>0.042</td>
<td>-0.041</td>
<td>0.014</td>
<td>0.333</td>
<td>-0.084</td>
</tr>
<tr>
<td>Feeling</td>
<td>(feels, touch)</td>
<td>0.097</td>
<td>-0.127</td>
<td>-0.236</td>
<td>0.244</td>
<td>0.005</td>
</tr>
<tr>
<td>Biological Processes</td>
<td>(eat, blood, pain)</td>
<td>-0.066</td>
<td>0.206</td>
<td>0.005</td>
<td>0.557</td>
<td>-0.239</td>
</tr>
<tr>
<td>Body</td>
<td>(check, hands, spit)</td>
<td>0.031</td>
<td>0.083</td>
<td>-0.079</td>
<td>0.122</td>
<td>-0.299</td>
</tr>
<tr>
<td>Health</td>
<td>(cicnic, flu, pill)</td>
<td>-0.277</td>
<td>0.164</td>
<td>0.059</td>
<td>-0.012</td>
<td>-0.004</td>
</tr>
<tr>
<td>Ingestion</td>
<td>(dish, eat, pizza)</td>
<td>-0.108</td>
<td>0.247</td>
<td>0.013</td>
<td>-0.058</td>
<td>-0.202</td>
</tr>
<tr>
<td>Work</td>
<td>(job, major, serum)</td>
<td>0.231</td>
<td>-0.096</td>
<td>0.330</td>
<td>-0.125</td>
<td>0.426</td>
</tr>
<tr>
<td>Achievement</td>
<td>(earn, hero, win)</td>
<td>-0.005</td>
<td>-0.240</td>
<td>-0.198</td>
<td>0.070</td>
<td>0.008</td>
</tr>
<tr>
<td>Money</td>
<td>(audit, cash, owe)</td>
<td>-0.063</td>
<td>-0.259</td>
<td>0.099</td>
<td>-0.074</td>
<td>0.222</td>
</tr>
<tr>
<td>Religion</td>
<td>(altar, church, mosque)</td>
<td>-0.152</td>
<td>-0.151</td>
<td>-0.025</td>
<td>0.383</td>
<td>-0.073</td>
</tr>
<tr>
<td>Death</td>
<td>(bury, coffin, kill)</td>
<td>-0.001</td>
<td>0.064</td>
<td>-0.332</td>
<td>-0.054</td>
<td>0.120</td>
</tr>
<tr>
<td>Fillers</td>
<td>(blah, mean, youknow)</td>
<td>0.099</td>
<td>-0.186</td>
<td>-0.272</td>
<td>0.080</td>
<td>0.120</td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commas</td>
<td></td>
<td>0.148</td>
<td>0.080</td>
<td>-0.24</td>
<td>0.155</td>
<td>0.170</td>
</tr>
<tr>
<td>Colons</td>
<td></td>
<td>-0.216</td>
<td>-0.153</td>
<td>0.322</td>
<td>-0.015</td>
<td>-0.142</td>
</tr>
<tr>
<td>Question Marks</td>
<td></td>
<td>0.263</td>
<td>-0.050</td>
<td>0.024</td>
<td>0.153</td>
<td>-0.114</td>
</tr>
<tr>
<td>Exclamation Marks</td>
<td></td>
<td>-0.021</td>
<td>-0.025</td>
<td>0.260</td>
<td>0.317</td>
<td>-0.295</td>
</tr>
<tr>
<td>Parentheses</td>
<td></td>
<td>-0.254</td>
<td>-0.048</td>
<td>-0.084</td>
<td>0.133</td>
<td>-0.302</td>
</tr>
<tr>
<td>Non-LIWC Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL Sentiment</td>
<td></td>
<td>0.177</td>
<td>-0.130</td>
<td>-0.084</td>
<td>-0.197</td>
<td>0.268</td>
</tr>
<tr>
<td>Number of Hashtags</td>
<td></td>
<td>0.066</td>
<td>-0.044</td>
<td>-0.030</td>
<td>0.217</td>
<td>-0.268</td>
</tr>
<tr>
<td>Words per tweet</td>
<td></td>
<td>0.285</td>
<td>-0.065</td>
<td>-0.144</td>
<td>0.031</td>
<td>0.200</td>
</tr>
<tr>
<td>Links per tweet</td>
<td></td>
<td>-0.061</td>
<td>0.081</td>
<td>0.256</td>
<td>-0.054</td>
<td>0.064</td>
</tr>
</tbody>
</table>
Figure 8: Expected values and standard deviation of personality characteristics, normalized on a 0–1 scale.

CHAPTER 4: Analysis

This chapter presents the results of the research. Analysis of those results shows that the mean value of each character trait matched those from the random selection of Twitter users in the study by Golbeck et al. [10]. The high level conclusion is that, while the personality traits of Navy Twitter users can be determined based on their Twitter activity, that information is insufficient as the sole predictor of who will be a good fit for Navy service.

4.1 Results

Using the formulas as described in Section 3.3, every user’s level of each of the five personality traits was calculated. Figure 9 shows a boxplot for each of the character traits, with the colored area representing the values between the first and third quartiles, the outer horizontal lines representing the minimum and maximum values, the horizontal line in the colored area indicating the median value, and the small circles representing the outliers, except those discussed in Subsection 4.2.

The formula used to calculate each user’s character traits based on their LIWC textual analysis and the means and correlations from [10] resulted in the mean of the sample of Navy users matching the mean of the sample from [10]. As a result, comparing the means of the Navy population to the wider Twitter population is not possible. However, some other useful information can be derived from other statistics.

4.1.1 Character Trait Distributions

Although the mean value of each character trait matched that given in Golbeck et al., the standard deviations of the traits of the Navy users were generally lower than those given in that work [10]. The standard deviations of each of the traits in the Five Factor Model from Golbeck et al., and this work are displayed in Table 3.

Conscientiousness was the only trait that had essentially the same standard deviation between the earlier work and this research; the other four traits had significantly lower standard
Figure 9: Boxplot showing the results of each of the Five Factors, with outliers trimmed.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Population</td>
<td>0.162</td>
<td>0.176</td>
<td>0.190</td>
<td>0.224</td>
<td>0.147</td>
</tr>
<tr>
<td>Navy Population</td>
<td>0.090</td>
<td>0.178</td>
<td>0.115</td>
<td>0.144</td>
<td>0.123</td>
</tr>
</tbody>
</table>

Table 3: Standard deviation of character traits for Navy personnel and earlier research.

... deviations. This shows that the Navy population has a more homogeneous personality makeup than the random selection of Twitter users from Golbeck et al.

As expected in a population, each trait displays a normal distribution. The Agreeableness values, as shown in Figure 10, have a narrow, sharp peak at the average, reflecting the low standard deviation seen in Table 3. This indicates that most of the Navy Twitter users had about average levels of Agreeableness. Conscientiousness, which had the highest standard deviation of the traits, exhibits a smoother curve with heavier tails at either end, as shown in
Figure 11. This finding was surprising because the work by Cooper and Pervin [6] showed that Conscientiousness is the trait most closely linked to job performance, and the sample population is of well-performing Navy personnel.

![Distribution of Agreeableness Values](image)

**Figure 10: Density graph of Agreeableness values.**

The distribution of Extroversion values is wider than that of Agreeableness, but narrower than Conscientiousness, with small peaks in the low end of the tail, indicating that, although the majority of the sample of Navy personnel have about average levels of Extroversion, there are a significant number that have very low to low Extroversion. As shown by DeJong et al. [15], people with higher levels of Extroversion adjust more easily to the military lifestyle; this work shows that those with lower levels of Extroversion can still perform well in a military lifestyle.

The density graph of Neuroticism values, as shown in Figure 13, displays a significant side peak below the overall average. This is consistent with the findings of DeJong et al. [15] that lower levels of Neuroticism correlate with ease of adjustment to the military lifestyle. The density graph of Openness to Experience, Figure 14, displays similar characteristics as the Extroversion graph but with a small rise above the average, very near the maximum possible value of 1. These users with very high levels of Openness to Experience are again consistent with DeJong et al. [15].
4.1.2 Non-language Correlations
Correlations between each trait and the non-language data points as defined in Section 3.3.1 were calculated; the full results are displayed in Table 4. The shaded cells indicate those correlation coefficients which were significantly different from 0, where p < 0.05. There was no strong correlation seen between any of the non-language data and a user’s level of each of the character traits; replies per tweet had a moderate correlation with both Agreeableness and Extroversion.

4.2 Calculation Anomalies
Although the measure of each trait should be between zero and one, each of the traits had a few users whose results were outside of that bounding, with values either below zero or above one. These errors occurred due to a disproportionately high value for one or more of the LIWC categories used to calculate the trait value. These users generally had a very small input size; of the 30 users who had at least one trait outside of the expected range, 25 had fewer than 200 words in their Twitter sample. These values outside of the expected range were included in all statistical calculations but are not displayed on any of the plots in this chapter.
Figure 12: Density graph of Extroversion values.

Figure 13: Density graph of Neuroticism values.
Figure 14: Density graph of Openness to Experience values.

![Distribution of Openness to Experience Values](image)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Followers</td>
<td>-0.033</td>
<td>0.042</td>
<td>0.017</td>
<td>0.021</td>
<td>-0.001</td>
</tr>
<tr>
<td>Following</td>
<td>-0.009</td>
<td>0.016</td>
<td>-0.057</td>
<td>0.022</td>
<td>-0.050</td>
</tr>
<tr>
<td>Favorites</td>
<td>0.004</td>
<td>-0.010</td>
<td>-0.005</td>
<td>-0.038</td>
<td>0.007</td>
</tr>
<tr>
<td>Tweets</td>
<td>0.108</td>
<td>-0.012</td>
<td>0.016</td>
<td>-0.018</td>
<td>-0.086</td>
</tr>
<tr>
<td>Retweets</td>
<td>0.051</td>
<td>-0.021</td>
<td>0.005</td>
<td>-0.043</td>
<td>-0.006</td>
</tr>
<tr>
<td>Replies</td>
<td>0.115</td>
<td>0.002</td>
<td>0.110</td>
<td>-0.032</td>
<td>-0.031</td>
</tr>
<tr>
<td>Hashtags</td>
<td>0.007</td>
<td>0.032</td>
<td>-0.037</td>
<td>-0.032</td>
<td>-0.071</td>
</tr>
<tr>
<td>Media</td>
<td>0.042</td>
<td>-0.031</td>
<td>-0.096</td>
<td>-0.061</td>
<td>-0.107</td>
</tr>
<tr>
<td>Words per tweet</td>
<td>0.009</td>
<td>0.126</td>
<td>-0.015</td>
<td>-0.075</td>
<td>-0.067</td>
</tr>
<tr>
<td>Retweets per tweet</td>
<td>0.053</td>
<td>-0.021</td>
<td>-0.013</td>
<td>-0.034</td>
<td>-0.001</td>
</tr>
<tr>
<td>Replies per tweet</td>
<td>0.228</td>
<td>0.060</td>
<td>0.279</td>
<td>-0.094</td>
<td>-0.025</td>
</tr>
<tr>
<td>Hashtags per tweet</td>
<td>-0.114</td>
<td>0.038</td>
<td>-0.040</td>
<td>-0.024</td>
<td>-0.037</td>
</tr>
<tr>
<td>Media per tweet</td>
<td>0.026</td>
<td>-0.043</td>
<td>-0.152</td>
<td>-0.060</td>
<td>-0.132</td>
</tr>
<tr>
<td>Followers/Following</td>
<td>-0.029</td>
<td>0.043</td>
<td>0.018</td>
<td>0.019</td>
<td>-0.005</td>
</tr>
</tbody>
</table>

Table 4: Correlation between character traits and non-language data. Shaded cells indicate correlation coefficients that are significantly different from 0, where p < 0.05.
After calculating the personality traits for each user, all of the data was then stored in a graph database to allow for easier access to the data and more complex data analysis. This chapter explains the model used to represent the data in a graph database, and identifies some of the questions that can be answered by querying the data.

5.1 Graph Database Model

The graph database program used to store the data from this research is Neo4j. As discussed in Section 2.3, Neo4j stores data as either a node, a relationship, or a property of a node or relationship. The overall model used to store this data is shown in Figure 15 and explained in more detail in this section.

5.1.1 Labels

The following labels were used to group the node data:

- User: a node to represent a user
- Tweet: a node to represent a tweet
- Hashtag: a node to represent a hashtag
- Location: a node to represent a latitude and longitude
- Characteristic: a node to represent one of the five personality characteristics in the Five Factor Model as described in Section 2.1
- Timeline: a single node used to organize the date and time references as described in Section 5.1.5
- Year: a node to represent a year from 2008 to 2015
- Month: a node to represent each of the months of a year
- Day: a node to represent each day of a month
5.1.2 Properties

Properties are additional data stored with a node or relationship; each instance of a node type may have any or all of these properties. A Location node has the properties of latitude and longitude. Each ACCOUNT_CREATED_ON and TWEETED_ON relationship has a property of time, and the relationship HAS_CHAR_TRAIT has a property defining where that User falls with that character trait on a scale from zero to one, using the calculations from Section 3.3. A User node has properties as listed in Table 5; a Tweet node has properties as listed in Table 6. The values of these properties come from Twitter as described in Section 2.2.1.

5.1.3 User Relationships

User nodes can be connected to other nodes in the following relationships:
<table>
<thead>
<tr>
<th>id</th>
<th>screen_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>default_profile</td>
</tr>
<tr>
<td>default_profile_image</td>
<td>description</td>
</tr>
<tr>
<td>favourites_count</td>
<td>followers_count</td>
</tr>
<tr>
<td>friends_count</td>
<td>geo_enabled</td>
</tr>
<tr>
<td>lang</td>
<td>listed_count</td>
</tr>
<tr>
<td>location</td>
<td>protected</td>
</tr>
<tr>
<td>statuses_count</td>
<td>time_zone</td>
</tr>
<tr>
<td>url</td>
<td>verified</td>
</tr>
</tbody>
</table>

Table 5: List of properties of a User node.

<table>
<thead>
<tr>
<th>id</th>
<th>screen_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>favorite_count</td>
<td>in_reply_to_screen_name</td>
</tr>
<tr>
<td>in_reply_to_status_id</td>
<td>in_reply_to_user_id</td>
</tr>
<tr>
<td>lang</td>
<td>sensitive</td>
</tr>
<tr>
<td>retweet_count</td>
<td>text</td>
</tr>
<tr>
<td>type</td>
<td>url</td>
</tr>
</tbody>
</table>

Table 6: List of properties of a Tweet node.

- User **TWEETED** a Tweet
- User **HAS_CHAR_TRAIT** Characteristic
- User **ACCOUNT_CREATED_ON** a Day
- User’s **CURRENT** Tweet
- Tweet **MENTSIONS** a User

Figure 16 provides a graphical representation of all of the possible relationships for a User node.

### 5.1.4 Tweet Relationships

Tweet nodes can be connected to other nodes in the following relationships:

- Tweet was **TWEETED_ON** a Day
- User **TWEETED** a Tweet
- User’s **CURRENT** Tweet
Figure 16: Neo4j output showing all of the possible ways a User can be connected to another node.

- Tweet MENTIONS a User
- Tweet CONTAINS a Hashtag
- Tweet is connected to a User’s PREVIOUS Tweet
- Tweet is connected to a User’s NEXT Tweet
- Tweet RETWEETED another Tweet
- Tweet was in REPLY_TO another Tweet
- Tweet has a TWEET_LOCATION of Location
- Tweet CONTAINS a Hashtag

Figure 17 provides a graphical representation of the possible relationships for a Tweet node.

5.1.5 Time Relationships

Years, Months and Days nodes are connected to each other using the following relationships:

- User ACCOUNT_CREATED_ON a Day
- Tweet was TWEETED_ON a Day
- Timeline contains the YEAR Year
- Year HAS a Month
- Month HAS a Day

Each user profile and tweet has a date and time associated with its creation. The dates were built using a timeline tree, as depicted in Figures 18–21. Each year represented in the data,
from 2008 to 2015, has a Year node, and each Year node has its own set of 12 Month nodes. Each Month node has its own set of Day nodes. Users and Tweets are linked to Days with an ACCOUNT_CREATED_ON or TWEETED_ON relationship with the time of creation stored as a property of the relationship.

5.2 Querying the Data
Once all the data has been imported into the database, it can be queried to find answers about the data. Queries are written using the language Cypher as described in Section 2.3.

One simple query would be to identify which users have a high level of Conscientiousness, which has a strong correlation with job performance [6]. The Cypher query to answer that question is:

```
MATCH (u:User)-[r:HAS_CHAR_TRAIT]->(:Characteristic
{name:"Conscientiousness"})
WHERE r.level > 0.7
RETURN u
```

Researchers also showed that successful adjustment to the military lifestyle was correlated
with higher levels of Extroversion and Openness to Experience and lower levels of Neuroticism [15]. The Cypher query to search the database to identify how many of the Navy users meet that standard is:

```
MATCH (u:User)-[r:HAS_CHAR_TRAIT]->(:Characteristic {name:"Extraversion"})
WHERE r.level > 0.7
MATCH (u)--[s:HAS_CHAR_TRAIT]->(:Characteristic {name:"Openness to Experience"})
WHERE s.level > 0.7
MATCH (u)--[t:HAS_CHAR_TRAIT]->(:Characteristic {name:"Openness to Experience"})
WHERE t.level < 0.4
RETURN count(u)
```

Although this research focused on personality traits, there are many more questions that can be asked about the data once it is in a database. One interesting question would be to identify
the users who have interacted with the official U.S. Navy Twitter account, @USNavy, by retweeting a tweet originally posted by the U.S. Navy account. The Cypher query to identify these users is:

```
MATCH (u:User {scr_name:"USNavy"})-[[:TWEETED]]->(t:Tweet)
MATCH (t)<-[:RETWEETED]-(v:Tweet)<-[:TWEETED]-(w:User)
RETURN w
```

Because a user can tag their tweets with a location, that information can be extracted provide a view of where Sailors are tweeting from. The Cypher query to identify which users are geotagging their tweets and all of the locations is:

```
MATCH (l:Location)<-[:TWEET_LOCATION]-(t:Tweet)
  <-[:TWEETED]-(u:User)
RETURN l, u
```

Storing the data in a database allows both more complicated queries related to the initial research question as well as a broader range of queries on the data.
CHAPTER 6: Conclusion and Future Work

This chapter presents the overall conclusions of this research as well as recommendations for future work in both the use of Twitter activity to determine personality and the use of personality information to determine fitness for Naval service.

6.1 Conclusions

This research was conducted to answer two questions:

• Can the personality characteristics of well-performing Navy personnel be determined based on their use of the Twitter social media platform?
• Can useful information be determined about a user’s personality and activity in order to differentiate Navy Twitter users from the general Twitter user population?

The finding of this research is that it is possible to determine the personality characteristics of Navy personnel based solely on textual analysis of their Twitter posts. With the exception of the few anomalies discussed in Section 4.2, a user’s level of each of the personality traits of the Five Factor Model was successfully calculated.

On the other hand, this research also discovered that determining a user’s personality does not provide enough useful information to differentiate between Navy users and non-Navy users. The method of calculating a user’s personality traits as explained in Section 3.3 did not permit the comparison of averages between the non-Navy population studied in [10] and the Navy population used in this research, and little other useful information could be determined from the statistics of the Navy population. There was also almost no correlation between a user’s personality and their Twitter activity, with only one non-language factor having a moderate level of correlation with a personality characteristic.

Although there was some useful information in the results, the primary conclusions of this research is that using textual analysis and the correlation data from [10] is insufficient to identify specific traits that make Navy personnel stand out on Twitter.
6.2 Future Work

Despite the finding that this method of simple textual analysis is insufficient to use as the basis of a model for identifying future Navy recruits, developing a social media-based model for Navy recruitment is still an important research area. There are several ways that further research in this area can be continued.

The first recommendation for future work is to have each of the users in the test population take a previously validated personality test to determine his or her levels of each personality characteristic. This implementation, although more difficult and resource-intensive than the method used in this research, would provide a stronger basis for comparison against the findings in [10] without the weakness of having to use the mean and standard deviation from that work. This might eliminate the problem where the two populations have the exact same mean for each of the traits, which would allow more useful information to be determined from the means. This method would also allow the use of a personality model other than the Five Factor Model.

Another recommendation for future work in this area is to use Twitter data from a population of well-performing Navy users, poorly-performing Navy users, and a similar group of non-Navy users in order to build a classifier that can determine which of these categories a user belongs to. This classifier could then be used to determine whether another user should be in the Navy—that is, if the user shows similar characteristics to those who have succeeded in the Navy. This classifier would be a vital part of a social media-based recruiting tool.

Further research should also be conducted to determine what the best personality characteristics are for different jobs in the Navy. For example, it seems likely that the personalities of those who succeed in jobs such as Information Systems Technician, Steelworker, and Commanding Officer of a ship are quite different. Having the information about different jobs would enable recruiters to target potential recruits with exactly the characteristics needed for the open positions.

Beyond just Twitter or other social media platforms, research should be conducted into the creation and validation of a personality-based assessment for entrance into the Navy or for future promotions. The U.S. Army has been administering the Tailored Adaptive Personality Assessment System (TAPAS) to new recruits at Military Entrance Processing
Stations since 2009, but not actually using it to screen out recruits [21]. Studies following
the tested recruits have shown that those who had poor scores on TAPAS have had generally
had poor performance in the Army, thus validating its results [21]. The U.S. Navy should
begin to use this test to collect data on its validity for Navy personnel before using it as a
general screening method at recruiting stations in order to identify those people who are
not a good fit for Naval service.
List of References


Initial Distribution List

1. Defense Technical Information Center  
   Ft. Belvoir, Virginia

2. Dudley Knox Library  
   Naval Postgraduate School  
   Monterey, California