



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Biosurveillance Using Clinical Diagnoses and Social Media Indicators in Military Populations

February 2017

CD Corley
S Volkova
J Rounds
LE Charles

JJ Harrison
J Mendoza
K Han

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <<http://www.ntis.gov/about/form.aspx>>
Online ordering: <http://www.ntis.gov>



This document was printed on recycled paper.

(8/2010)

Acknowledgments

This work was supported by the Defense Threat Reduction Agency under contract CB10082 to Pacific Northwest National Laboratory. The authors thank Commander Jean-Paul Chretien, Aaron Kite-Powell, and Vivek Khatri, who were at the Armed Forces Health Surveillance Branch, Defense Health Agency, for helpful discussions on the research and experimental design of this contract.

Acronyms and Abbreviations

AFHSC	Armed Forces Health Surveillance Center
BSVE	Biosurveillance Ecosystem
CDC	U.S. Centers for Disease Control
ETL	Extract Transform Load
EWMA	Exponentially Weighted Moving Average
GTSEDA	Generalized Time-Series Exploratory Data Analysis
ICD-9	International Classification of Disease, 9th edition
ILI	Influenza-like illness
ILINet	U.S. Outpatient Influenza-like Illness Surveillance Network
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short-Term Memory
MMWR	Morbidity and Mortality Weekly Report
NNDSS	National Notifiable Diseases Surveillance System (CDC)
SODA	Socrata Open Data API
STL	Seasonal Trend Loess
VADER	Valence Aware Dictionary and sEntiment Reasoner

Contents

Executive Summary	Error! Bookmark not defined.
Acknowledgments.....	iii
Acronyms and Abbreviations	iv
1.0 Project Overview	1.1
1.1 Objective	1.1
1.2 Introduction	1.1
1.3 Methods and Results	1.1
1.4 Conclusion.....	1.2
2.0 Generalized Time Series Exploratory Data Analysis Application	2.1
2.1 Introduction	2.1
2.1.1 Data Sources.....	2.2
2.2 User Interface	2.3
2.2.1 Query Tab.....	2.4
2.2.2 Select Tab.....	2.4
2.2.3 Summary Tab	2.6
2.2.4 Clustering Tab	2.7
2.2.5 Within-Location Clustering Tab	2.10
2.2.6 Alarms Tab.....	2.14
2.3 Future Work and Discussion	2.15
2.3.1 Data Sources.....	2.15
2.3.2 Automation.....	2.16
2.3.3 Modularization	2.16
2.3.4 Further Output Integration	2.17
3.0 Military Biosurveillance: Studying Military Community Health, Well-being, and Discourse through the Social Media Lens	3.1
3.1 Research Questions	3.1
3.2 Background and Related Work	3.2
3.2.1 Characteristics of the U.S. Military Population	3.2
3.2.2 Studies on Military Populations	3.2
3.2.3 Understanding Populations through Social Media.....	3.2
3.3 Data	3.3
3.3.1 Data Anonymization	3.3
3.3.2 Sampling Military Users on Twitter.....	3.3
3.4 Analysis and Results	3.4
3.4.1 RQ1: Differences in Social Media Activities of Military vs. Control.....	3.4
3.4.2 RQ2: Differences in Language Use between Military and Control	3.5
3.4.3 RQ3: Trends of Sentiment for Military and Control.....	3.1

3.4.4	RQ4: Topic Variations between Military and Control	3.2
3.4.5	RQ5: Health-related Discourse between Military and Control	3.4
3.5	Discussion	3.5
3.5.1	Implications for Military Social Life.....	3.5
3.5.2	Implications for Military and Public Health.....	3.5
3.5.3	Limitations and Future Work	3.6
3.6	Conclusion.....	3.6
3.7	Acknowledgment	Error! Bookmark not defined.
4.0	Military Biosurveillance: Predicting Influenza Dynamics with Neural Networks Using Signals from Social Media.....	4.1
4.1	Motivation	4.1
4.2	Approach.....	4.1
4.3	Results	4.1
5.0	Chiron Computing Pipeline and Architecture	5.1
5.1	Server Architecture and Description	5.1
5.2	ETL Pipeline	5.2
5.2.1	High-level ETL	5.2
5.2.2	Enrichments.....	5.3
6.0	Understanding Readers' Credibility Perceptions on Social Media Content: Case of Disease Outbreaks on Twitter	6.1
6.1	Introduction	6.1
6.2	Related Work	6.2
6.2.1	Understanding Credibility in Social Media.....	6.2
6.3	Study Goals	6.2
6.4	Study Design	6.3
6.5	Results	6.4
6.5.1	Survey Respondent Demographics	6.4
6.5.2	RQ1: Impact of Features on Perceived Credibility	6.4
6.5.3	RQ2: Variance in Perceived Credibility.....	6.5
6.5.4	RQ3: Reader and Author Factors on Perceived Credibility	6.6
6.6	Discussion and Conclusion	6.8
7.0	Software Delivered	7.1
8.0	Publications	8.1
9.0	References	9.1

Figures

Figure 2.1. The Larger Vision for GTSEDA. The use of geolocated time-series data that are common in syndrome surveillance is the primary focus.....	2.1
Figure 2.2. Primary Information Display in App. AFHSB hospitals are displayed. Map values are current Exponentially Weight Moving Average values for Influenza-Like Illness for ages 0 to 4. This map and time-series plot is the primary display consistent across all tab types within the app. Prev and Next buttons move threw locations, but the map can also be clicked to select locations.	2.3
Figure 2.3. Below the Primary Information Display (Figure 2.2), a Tab Set for Navigation of GTSEDA Capabilities Is Provided. The Query tab is selected in this figure.	2.4
Figure 2.4. Select Tab. Time-series can be selected by knowledge base properties and transformed for further operations.	2.5
Figure 2.5. Ft. Sill Rates of Influenza-like Illness in Children 0 to 4. This time-series is not available on BSVE, but it illustrates the STL decomposition of a time-series within the summary panel. (Figure 2.6).	2.6
Figure 2.6. Seasonal Trend Loess Decomposition of a Ft. Sill Time Series. The upper panel shows the cyclic seasonal component observed in this time-series. The middle panel shows the trend, and unusually in 2009 there is an upward trend peaking in August 2009. The bottom panel is the residual left over after subtracting the seasonal component and the trend.	2.7
Figure 2.7. For Most Clustering Operations, Users Should First Apply the “Scale” and “Center” Transformations before Clustering. If the time-series are not scaled and centered, the clusters are valid, but they will be on the magnitude of the observations.	2.8
Figure 2.8. UI for Clustering across Locations	2.8
Figure 2.9. Clustering AFHSB Hospitals by ILI for Ages 0-4 in the 2012 to 2013 Season	2.9
Figure 2.10. Cluster Means. The orange cluster (Midwest and Pacific Northwest) is different from the other clusters because it had its largest peak for ILI illness for ages 0 to 4 in late February. Selecting each location individually confirms this fact.	2.10
Figure 2.11. Cluster Means for Within-Location Clustering. Note that there are two distinct types of diseases in New York: those that always peak in the summer within the period of observation and those time-series that do not.	2.11
Figure 2.12. Dendrogram of Within-Location Clustering. Tracing diseases from bottom to top can help understand how clusters were merged.	2.12
Figure 2.13. Distance Matrix as a Heatmap Visualized within the App. Distances over 86 units have been filtered out (threshold selected interactively within the UI).	2.12
Figure 2.14. The Information of Figure 2.12 Re-interpreted as a Graph. Here edges are allowed between diseases when the heatmap Distance matrix distances are under a user-selected threshold. We can see that the seasonal bifurcation of New York diseases is preserved in the edges between diseases....	2.13
Figure 3.1. Monthly Trend of Positive and Negative Sentiment Scores.....	3.2
Figure 3.2. A: Distribution of Topics Based on Tweets for Military and Control Populations. B: Distribution of topics based on tweets for military and control populations grouped by geography. C: Distribution of topics based on tweets for military and control populations grouped by military service types. Colored area: Military population; non-colored area: Control population.....	3.3
Figure 5.1. The ETL Pipeline	5.2
Figure 5.2. Enrichments for the Social Data Analytics Pipeline.....	5.3

Figure 6.1. Procedure of the Survey. After answering personal information, respondents were asked to complete four types of tasks. The same 16 tweets were used in Tasks 3 and 4, and the corresponding 16 tweet pages were randomly assigned to survey respondents.	6.3
Figure 6.2. Example of the Tweet Page Used in Task 4. Four author factors (two conditions per factor) were measured. We have 200,000 samples of users who posted tweets related to disease outbreaks, and High and Low conditions in Tweet Attention and Author Engagement were decided based on the medians of each case from our samples. Note the High and Low conditions for each factors.....	6.4
Figure 6.3. Impacts of Author-based Features on Credibility Assessment.....	6.5
Figure 6.4. Impacts of Tweet-based Features on Credibility Assessment	6.5
Figure 6.5. Difference in Credibility Assessments between Task 2 and Task 3 for the Same 16 Tweets. Overall, respondents' credibility judgments were significantly influenced by other factors ($p < 0.05$).	6.6
Figure 6.6. Difference in Credibility Ratings in Task 4 between Two Conditions for Each Author Factor. All factors showed significant differences ($p < 0.05$). Note that Low in the author bio refers to no author description, and Low in the other tweets refers to non-professional tweets.	6.7
Figure 6.7. Difference in Credibility Ratings between Task 3 and 4 (all $p < .05$). Note that Low in the author bio refers to no author description, and Low in the other tweets refers to non-professional tweets.	6.7

Tables

1.1 Differences in Twitter Health-related Terminology between Military and Non-military Populations	1.2
2.1 Data A Reprocessed prior to the User Seeing It into a Table with the Following Schema	2.3
3.1 Military Locations $L_1 \dots L_6$ and the Corresponding Number of Users Sampled for Both Military and Control Populations Together. The total number of users sampled across six locations is 10,814. ...	3.3
3.2 Example Keywords Used to Identify Military Users	3.4
3.3 Comparing Mean Values for User Activities and Online Behavior across Military vs. Control Populations ($p\text{-value} \leq 0.001^{***}$, $p\text{-value} \leq 0.01^{**}$).....	3.5
3.4 Differences in Linguistic Attributes between Military and Control Populations Measured Using LIWC. We only present linguistic categories which have the same directions across populations. $\Delta = (\mu_{mil} - \mu_{con}) \times 10^{-3}$ ($p\text{-value} \leq 0.001^{***}$, $p\text{-value} \leq 0.01^{**}$	7
3.5 Keywords Specific to Each Military and Control Sample, Extracted Using SAGE (Eisenstein et al. 2011)	3.1
3.6 The Example of Health Category Keywords. A * indicates a regular expression; for example fever* indicates words that have a stem fever with difference suffixes such as fevers, feverish and fevered.	3.4
3.7 Comparing the Counts of Health Words for Military vs. Control Populations.....	3.4
5.1 The Three OpenStack Machine Sizes Used to Build the Chiron Data and Computing Architecture.	5.1
5.2 Servers Used in the Chiron BSVE Project.....	5.1
6.1 Summary of the Influence of Author and Reader Factors on Credibility Perception. All author factors showed significant influences, but the impact of author bio was incomparable to other factors.....	6.6

1.0 Project Overview

U.S. military influenza surveillance uses electronic reporting of clinical diagnoses to monitor health of military personnel and detect naturally occurring and bioterrorism-related epidemics. While accurate, these systems lack in timeliness. More recently, researchers have used novel data sources to detect influenza in real time and capture nontraditional populations. With data-mining techniques, military social media users are identified and influenza-related discourse is integrated along with medical data into a comprehensive disease model. By leveraging heterogeneous data streams and developing dashboard biosurveillance analytics, the researchers hope to increase the speed at which outbreaks are detected and provide accurate disease forecasting among military personnel.

1.1 Objective

The project objective is to integrate existing influenza surveillance data sources and social media data into an accurate and timely outbreak detection model embedded into dashboard biosurveillance analytics for the U.S. Department of Defense.

1.2 Introduction

Influenza-like illness (ILI) remains a significant public health burden to both the general public and the U.S. Department of Defense. Military personnel are especially susceptible to disease outbreaks owing to the often-crowded living quarters, substantial geographic movement, and physical stress placed upon them (Sueker et al. 2010). Currently, the military employs syndromic surveillance on electronic reporting of clinical diagnoses. While faster than traditional, biologically focused monitoring techniques, a recent study conducted by the Centers for Disease Control (CDC 2009) found that the military surveillance system proved inadequate at detecting outbreaks quickly enough. Recently, research has included novel data sources, like social media, to conduct disease detection in real time and capture communities not traditionally accounted for in current surveillance systems. Data-mining techniques are used to identify influenza-related social media posts and train a model against validated medical data (Fairchild 2014). By integrating social media data and a medical dataset of all ILI-related laboratory specimens and doctor visits for the entire military cohort, a more comprehensive model than presently exists for disease identification and transmission will be possible.

1.3 Methods and Results

For analyses, the Armed Forces Health Surveillance Center (AFHSC) provided about 1000 military health facilities' Defense Medical Surveillance System data, recorded between December 1999 and 2014. These data included laboratory results and medical clinical visits coded with an International Classification of Disease, 9th edition (ICD-9) code under the AFHSC's syndromic definition of ILI. Health facilities were mapped in ESRI ArcGIS with a 25-mile buffer. To determine specific locations of interest for historical Twitter data purchase and analyses, facilities within each buffer were condensed into a merged location and areas with substantial medical data, military populations, and social media usage were targeted. From this analysis, twenty-five U.S. and six international condensed locations were chosen as study sites. Three additional non-military locations, based on comparative attributes, were identified as control sites. Geo-tagged tweets, from November 2011 to June 2015, were purchased within a 25-mile radius of the centroid for each of the thirty-one identified locations of interest.

Descriptive summary statistics for each location, time series analyses, and correlation studies of ICD-9 codes and laboratory data against regional CDC U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and city-level Google Flu Trends were conducted. Social media analytics on military and non-military tweets identified differences in Twitter discourse between the two cohorts, including common language, sentiment and health-related topics (Table 1.1).

Table 1.1. Differences in Twitter Health-related Terminology between Military and Non-military Populations

Category	Mean (Military)	Mean (Control)	T-statistic	P-value
Self-related health experience	0.0037	0.0031	3.907	9.74E-05
ILI-specific symptoms	0.0008	0.0008	0.261	7.94E-01
Disease names and terms	0.0012	0.0012	0.668	5.04E-01
Entities	0.0012	0.0012	0.559	5.77E-01
Parts of body and related	0.0003	0.0003	-1.216	2.24E-01
Non-ILI specific symptoms	0.0006	0.0006	-0.382	7.01E-01

1.4 Conclusion

Twitter flu-related discourse from military members and electronic medical data will be incorporated into a robust outbreak detection model. This model will continually ingest new health and social media data to nowcast and forecast influenza activity on military bases. A user-friendly application will provide military analysts with tools required to allocate resources efficiently and effectively.

2.0 Generalized Time Series Exploratory Data Analysis Application

The Generalized Time-Series Exploratory Data Application is deployed in the Biosurveillance Ecosystem (BSVE), and it provides interactive exploration and analysis of syndrome data sources such as the CDC National Notifiable Diseases Surveillance System (NNDSS) Table II data. This app provides exploratory data analysis and alerting capabilities to analysts. These capabilities include Seasonal Trend Loess time-series decompositions, hierarchical geo-location clustering, and alerting.

2.1 Introduction

Syndromic surveillance systems are an important part of the larger topic of biosurveillance systems deployed to protect the interest of the United States and partners. Typically, such systems generate aggregate geo-located time-series for many related diseases. Anomaly detectors, alerting algorithms, and other information generating systems analyze these data to create information important to an analyst.

The Generalized Time-Series Exploratory Data Analysis application (GTSEDA) is an app developed under contract with PNNL deployed on the BSVE. The purpose of GTSEDA is to address some common tasks and operations one might wish to perform on weekly geo-located syndrome time-series, but GTSEDA allows this to be done entirely from within the BSVE. In addition, emphasis was placed on allowing analysts to explore geolocated time-series, so that they can trust any inference made by the system. At times, this ability to explore the data came at the cost of automation, a topic discussed Section 2.4.

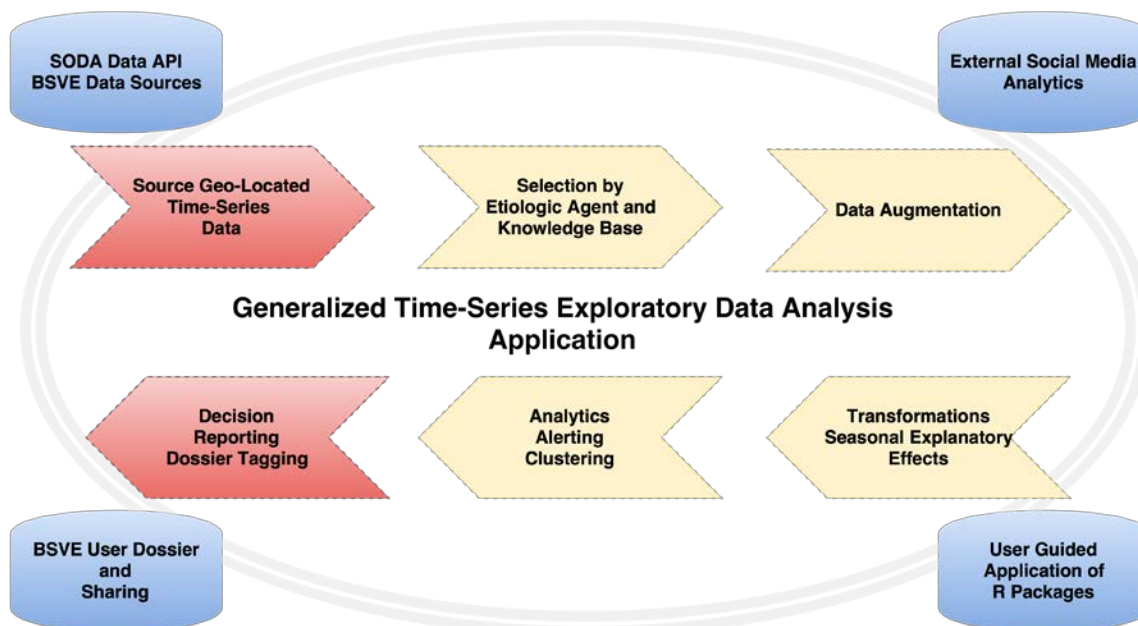


Figure 2.1. The Larger Vision for GTSEDA. The use of geolocated time-series data that are common in syndrome surveillance is the primary focus.

2.1.1 Data Sources

We introduce the data source experience within the application.

2.1.1.1 Background and Technical Remarks

Data sources were a challenging aspect of the BSVE experience to manage. A repeating aggregate geo-located time-series has a pretty clear structured data schema, but at the time the app was developed, there was not an efficient common schema within the BSVE. Furthermore, the existing data sources used by GTSEDA are not efficiently queried in its current schema or MongoDB format.

The approach taken for GTSEDA was to reshape selected BSVE data into a new schema and place that data onto the BSVE Postgres servers in an application table. Redeployment to BSVE Postgres was done in accordance with the BSVE documentation guide, but this plan basically forced two things upon GTSEDA: a CRON-like task would have to operate externally to BSVE to constantly update the new schema tables, and what tables a user could use would be dictated by the developers and sponsors of GTSEDA.

2.1.1.2 SODA dataSource and NNDSS

The primary source of data used within GTSEDA originally comes from the CDC. The NNDSS is a longstanding aggregate syndrome surveillance publication of the CDC. It consists of three tables; *Table I: Infrequently reported notifiable diseases*; *Table II* is divided into 14 parts with each part containing several weekly notifiable diseases (Morbidity and Mortality Weekly Report [MMWR] timestamps); and finally, there is *Table IV: Tuberculosis*. In the past, this list would have contained *Table III: U.S. Deaths in 122 Cities*, but it was discontinued in Week 40 of 2016 as part of a modernization effort. The data still exist and are reported by CDC via their own interface not provided to BSVE (see <https://wonder.cdc.gov/mmwr/mmwr morb.asp>).

CDC exports these tables to a broader community via an interface called Socrata Open Data API (SODA), which generally provides a JSON organized by MMWR week number for each location of two types: cities, states, or regions depending on context.

These details are hidden to GTSED users, but all four of these CDC NNDSS tables are present in BSVE as a dataSource. The dataSource has the type “SODA,” and then respective tables from the CDC are enumerated in the name. Unfortunately, JSON returned from accessing this SODA data source have many fields that are unnecessary to syndrome surveillance, so rather than use these data directly, we reprocess prior to deploying GTSEDA and in an ongoing background task for weekly updates.

One of the decisions made in pre-processing the SODA data to a new schema was to keep only the current week observation for each week. The SODA data are loaded with several extra fields that do not seem applicable the GTSEDA app.

2.1.1.3 GTSEDA Schema

The number of Postgres tables GTSEDA can interoperate with is not fixed, and conceptually distinct tables can be kept separate. The Postgres tables that GTSEDA uses are row-oriented data structures as shown in Table 2.1; however, the field names are not important. The GTSEDA app gets told which fields have the corresponding equivalent meaning in a JSON structure at startup. For future development, although the field names need not be the same, each field needs to have the same purpose, organization,

and type, that is, equivalent schema. The type field of Table 2.1 refers to different diseases; for example, it can take on values identifying unique diseases monitored in NNDSS.

Table 2.1. Data A Reprocessed prior to the User Seeing It into a Table with the Following Schema

Field Name	Purpose	Value Type
reporting_area or longname	Uniquely identify a geolocation	String
timestamp	datetime in Y/M/D format	String
series	Name of observation type or time-series	String
value	Value for observation at that time	Integer

2.2 User Interface

The user interface (UI) follows a natural progression of exploratory and analytic analysis for geolocated time series. The UI provides various tools to transform, summarize, and examine time-series, cluster time-series across geolocations and within geo-locations, and alert on unusually large quantities within time-series.

The UI is laid out vertically. At the top of the UI is a map widget and below that is the primary time-series currently being considered by the application (Figure 2.2). The map widget responds to “onclick” events and can be used to select a new geolocation for analysis or further consideration.

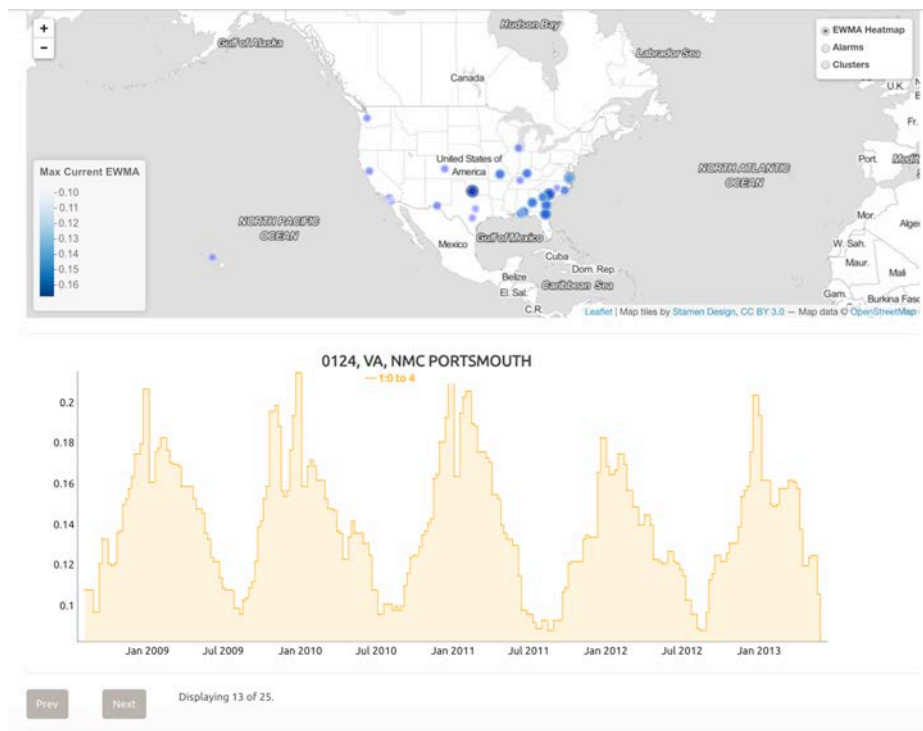


Figure 2.2. Primary Information Display in App. AFHSB hospitals are displayed. Map values are current Exponentially Weight Moving Average values for Influenza-Like Illness for ages 0 to 4. This map and time-series plot is the primary display consistent across all tab types within the

app. Prev and Next buttons move through locations, but the map can also be clicked to select locations.

2.2.1 Query Tab

Below the primary map and time-series is a tab set for app operation (Figure 2.3). Before using the app in any capacity, users should first select a data source and click **Query All Locations**.

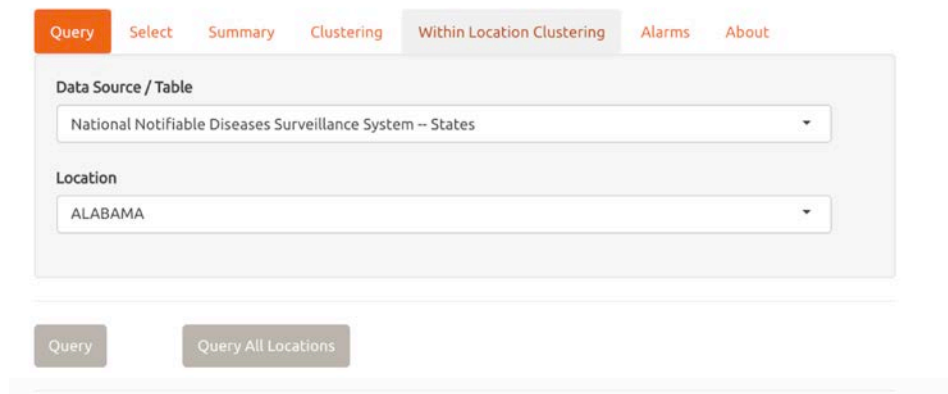
The screenshot shows a web application interface with a horizontal tab bar at the top. The tabs are labeled 'Query', 'Select', 'Summary', 'Clustering', 'Within Location Clustering', 'Alarms', and 'About'. The 'Query' tab is currently selected and highlighted in orange. Below the tabs is a form area with two dropdown menus. The first dropdown is labeled 'Data Source / Table' and has 'National Notifiable Diseases Surveillance System -- States' selected. The second dropdown is labeled 'Location' and has 'ALABAMA' selected. Below these dropdowns are two buttons: 'Query' and 'Query All Locations'. The 'Query All Locations' button is highlighted in a darker grey.

Figure 2.3. Below the Primary Information Display (Figure 2.2), a Tab Set for Navigation of GTSEDA Capabilities Is Provided. The Query tab is selected in this figure.

2.2.2 Select Tab

The Select tab contains more tools for customizing the currently selected data. Importantly, a knowledge base allows users to select disease time-series by interest or topic (Figure 2.4). In addition, users can transform the selected data to only include a time period of active interest.

Query **Select** Summary Clustering Within Location Clustering Alarms About

Series Options

Observe Time-Series For (variable)

babesiosis

Transformations

☒ Sum Selected ☐ Median Smoothing ☐ 1st Difference ☐ Scale ☐ Center ☐ Remove Seasonal Effect

☐ 52 Week Window

Measurement / Variable

value

Clear Selected

Knowledge Base Aggregation

Show 10 entries Search:

	Etiologic Agent	Transmission	Symptoms	Infect	Parasite
1	Babesiosis	tickborne	non-specific flu-like, anemia; asymptomatic if healthy	RBC	Protozoa
2	Campylobacteriosis	fecal-oral (contaminated food/water)	Gastrointestinal (GI)	Intestine	Bacteria

Figure 2.4. Select Tab. Time-series can be selected by knowledge base properties and transformed for further operations.

2.2.2.1 Methods

Several statistical methods can be applied to geo-located time-series. We briefly outline each and point to related documentation when available. For this document, we assume that time-series are on MMWR weekly observations.

2.2.2.2 Transformations

A variety of transformations are available to be applied to the time-series and are described below.

- **Sum:** If more than one time-series is selected from the knowledge base, then the counts of time-series can be summed prior to analysis via “sum” in Transformation.
- **Median Smoothing:** Especially noisy time-series may benefit from a median smoothing. In media smoothing, the observation at time t of a time-series V_t is replaced with the median of (V_{t-1}, V_t, V_{t+1}) so that the middle of a 3-week time span is used as the actual observation. This approach was found to be useful in analyzing especially noisy ILI at smaller military facilities.
- **1st Difference:** The time-series at time t is replaced by $D_t = V_t - V_{t-1}$. Some distributions are better modeled in this fashion. However, this feature is of limited usefulness for the MMWR analysis.

- **Center:** Observations are replaced by $V_t = V_t - \frac{\sum_{u=1}^{u=T} V_u}{T}$, where T is the number of observations in the time-series.
- **Scale:** Observations are replaced by $V_t = V_t / s.d(V)$. Checking scale and center together performs normalization of a time-series. Scale is useful for comparing unrelated disease counts and useful for measuring distance such as measurements employed in clustering.
- **Remove Seasonal Effect:** When the number of weeks in a series is greater than 104, it is mathematically feasible to remove a seasonal effect prior to analysis. This transformation does so with a Seasonal Trend Loess (STL) decomposition. This transformation estimates a seasonal component, as discussed in the Summary Tab section below, and subtracts it from the time-series for further analysis. More information about STL is available at <https://www.otexts.org/fpp/6/5>.
- **52-week window** is a helper function to clip a time-series to the last 52 weeks of observations. Users can also perform this manually at the bottom of the Select tab.

2.2.3 Summary Tab

The Summary tab quantifies the historic distribution of the current time-series. Common statistics such as the mean and standard deviation of the time-series are reported here. Also reported are the 25th and 75th quantiles of the distribution and the percentile of the current observation (Current Quantile). This information gives an immediate quantification of how extreme the present observation is relative to the time-series historic range.

If a time-series has more than two years of observations, an STL decomposition is also presented that breaks down the time-series into component parts. Figure 2.5 shows a typical STL time-series representation, and Figure 2.6 shows an STL representation in BSVE. This plot allows users to quickly assess whether a time-series appears to be drifting (trending) and to make distinct a true residual that might occur with a pandemic versus an expected annual seasonal component. More information about STL is available at <https://www.otexts.org/fpp/6/5>.



Figure 2.5. Ft. Sill Rates of Influenza-like Illness in Children 0 to 4. This time-series is not available on BSVE, but it illustrates the STL decomposition of a time-series within the summary panel. (Figure 2.6).

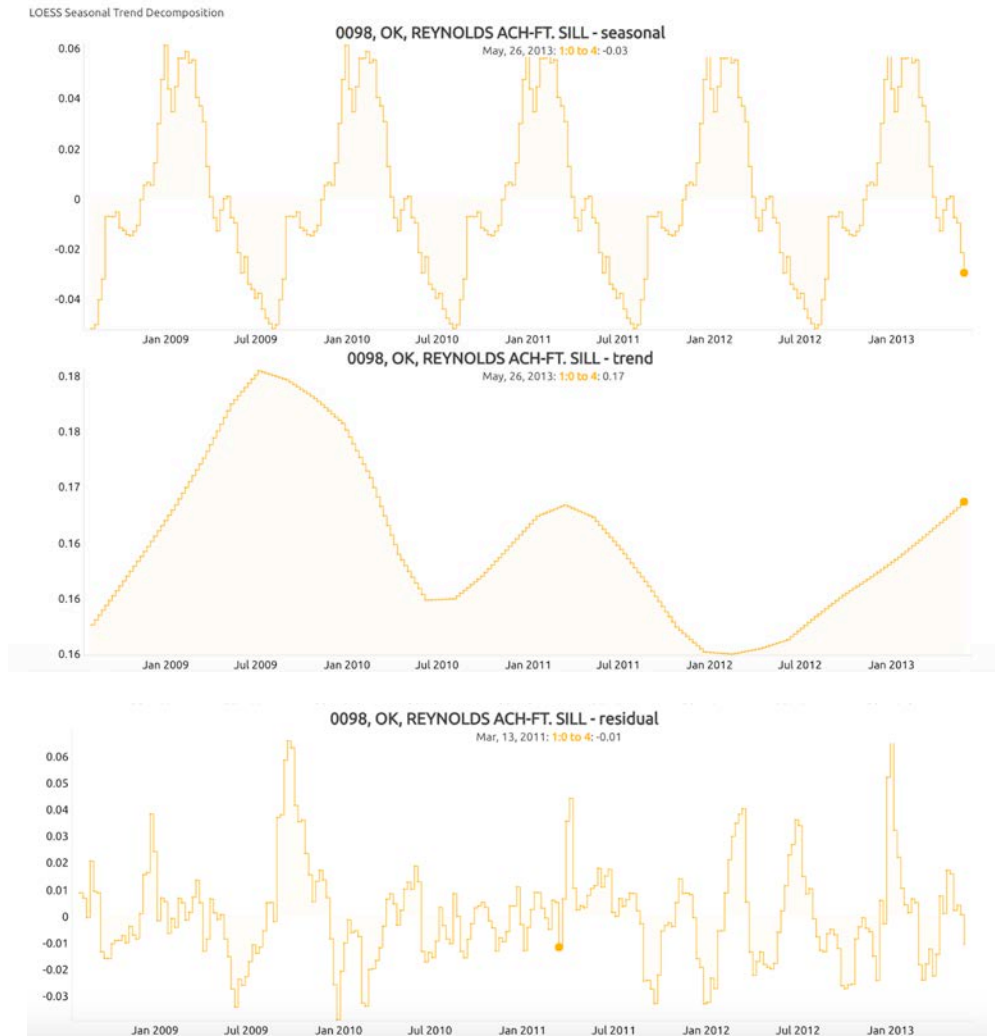


Figure 2.6. Seasonal Trend Loess Decomposition of a Ft. Sill Time Series. The upper panel shows the cyclic seasonal component observed in this time-series. The middle panel shows the trend, and unusually in 2009 there is an upward trend peaking in August 2009. The bottom panel is the residual left over after subtracting the seasonal component and the trend.

2.2.4 Clustering Tab

Clustering is perhaps the most complicated tab in the application. Jargon has been exposed directly from R functions. The general procedure is to pick a distance measure and then a hierarchical clustering method. The “Relation” selection in the Clustering tab is where a distance is picked: it was titled “relation” because we also allowed time-series to be related by correlation and partial correlation, but that may be changed back to Distance in a future version because the distance matrix must be calculated from any measurement before performing a clustering (see Figs. 2.7 and 2.8).

Series Options

Observe Time-Series For (variable)

giardiasis

Transformations

☐ Sum Selected ☐ Median Smoothing ☐ 1st Difference ☒ Scale ☒ Center ☐ Remove Seasonal Effect

☐ 52 Week Window

Measurement / Variable

value

Figure 2.7. For Most Clustering Operations, Users Should First Apply the “Scale” and “Center” Transformations before Clustering. If the time-series are not scaled and centered, the clusters are valid, but they will be on the magnitude of the observations.

Query Select Summary **Clustering** Within Location Clustering Alarms About

Relation

manhattan

Hierarchical Clustering Method

ward.D2

Number of Clusters

1 2 3 4 5 6 7

Cluster Locations

Figure 2.8. UI for Clustering across Locations

2.2.4.1 Relations/Distances

The clustering is across locations currently loaded into GTSEDA. Suppose we number each location $l \in \{1, \dots, L\}$, where L is the total number of locations loaded into GTSEDA. Further, we number the time-series MMWR weekly observations $t \in \{1, \dots, T\}$, where T is the total weekly observations. We calculate a distance between time-series v_r and v_s as $d(v_r, v_s)$ where the form of d depends on what is selected in Relation.

We summarize each selection here:

- **Manhattan** is $d(v_r, v_s) = \sum_t |v_{rt} - v_{st}|$.
- **Euclidean** is $d(v_r, v_s) = \sqrt{\sum_t (v_{rt} - v_{st})^2}$.
- **Correlation** is $d(v_r, v_s) = |1 - \rho_{rs}|$, where ρ_{rs} is the correlation between the time-series r and s .
- **Partial Correlation** is correlation between r and s after conditioning on the values of the other locations.

The remaining distance measures are more experimental in the context of the application. The above measures are emphasized because Manhattan distance seems to perform well in the context of proportional time-series (ILI rates); Euclidean distance seems to perform well in the context of count data; and finally, correlation is very easy to interpret. More information for the other distance options can be found at <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html>.

Distances are pairwise calculated between each and every location, so that there is a distance matrix that is L by L —that is the number of locations by the number of locations—upon which clustering is performed.

2.2.4.2 Hierarchical Clustering Method

Clustering is done through hierarchical clustering such that time-series are grouped according to rules, but the rules vary according to the method selected.

Options come directly from R, and recommended reading for these options is online in the details section of the underlying function at <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html>. Briefly, the important options are the following:

- Ward.D2. This algorithm precedes bottom up, joining clusters such that the variance within them is the lowest possible until all the locations are in one cluster. It generally discovers “compact spherical clusters.”
- Complete. This algorithm precedes bottom up, joining clusters such that the closest locations between clusters are as large as possible (so that clusters are forced apart).
- Single. This algorithm precedes bottom up, joining clusters such that the minimum distance between locations within two clusters is the smallest observed anywhere. It tends to find long “stringy” clusters and is related to a minimum spanning tree.

Ward.D2 makes the most geographically interpretable clusters, but we leave the other options active for users.

Finally, the user selects the number of clusters, which can be from 2 to 7, via slider.

2.2.4.3 Clustering Map Output

The most important output is a change to a map layer that colors locations by cluster membership. For example, Figure 2.9 is the clustering map output after selecting the 2011 to 2012 season of Figure 2.6 with default clustering options except that four clusters are used instead of five.

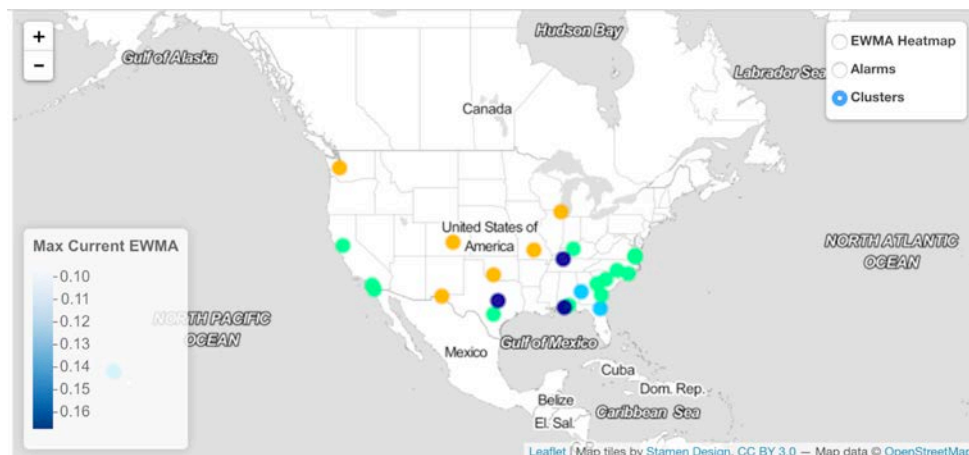


Figure 2.9. Clustering AFHSB Hospitals by ILI for Ages 0-4 in the 2012 to 2013 Season

Clustering Interpretation with Cluster Means

Each cluster has a mean over time for the time-series within the clusters. This takes the form of its own time-series, so that means change over time, and is particularly effective for the Ward.D2 method, which tries to minimize variances within clusters. In Figure 2.10, we see the cluster means for Figure 2.9. What is shown is that certain geolocations within the United States had their largest ILI rates for this demographic in March instead of December/January (Cluster 1—Orange). The effectiveness of clustering is that it helps guide the organization of geolocation into a common inference.

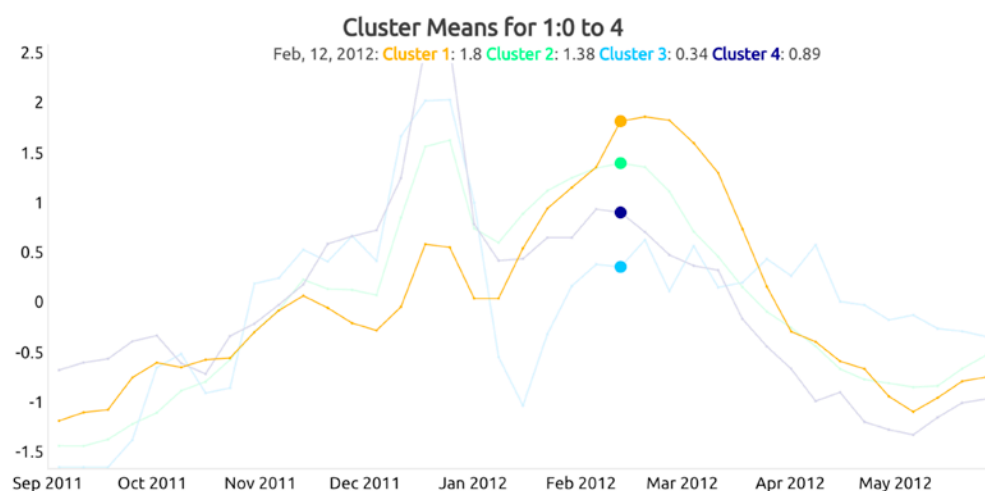


Figure 2.10. Cluster Means. The orange cluster (Midwest and Pacific Northwest) is different from the other clusters because it had its largest peak for ILI illness for ages 0 to 4 in late February. Selecting each location individually confirms this fact.

2.2.5 Within-Location Clustering Tab

Another type of clustering within the app groups diseases within a location. Unlike the Clustering tab, the things being organized within a group are not locations, but rather diseases. This can help create understanding of the seasonality of diseases within a location by grouping time-series. Its usefulness is limited to anomaly detection and scenarios where rapid assessment of data within a location is necessary. The difference between this tab's output and just examining the seasonality directly in the Summary tab is that here every disease known to a geo-location is examined at once.

For example, we loaded NNDSS Table II data for states available at the time of this writing within the app. On the Select tab, we selected “Center” and “Scale” transformations. We selected “New York” on the map UI, and then we selected the “Within Location Clustering.” We left the defaults on the Clustering tab and clicked Clusters. This produces output like that show in Figure 2.11 where it can be seen one cluster of diseases has more case counts in summer than in winter. This helps us identify typical seasons across many agents.

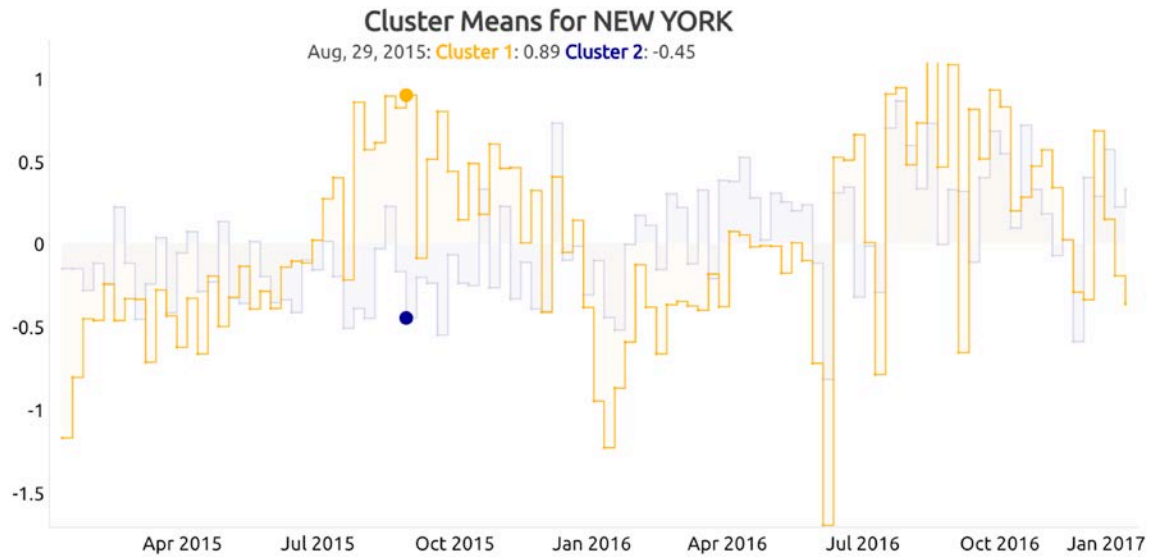


Figure 2.11. Cluster Means for Within-Location Clustering. Note that there are two distinct types of diseases in New York: those that always peak in the summer within the period of observation and those time-series that do not.

2.2.5.1 Clustering Interpretation with Cluster Dendrograms and Interactive Heatmaps

Cluster means alone, however, are not sufficient to interpret the WithinLocation Clusters, mostly because there is no associated map that indicates membership such as is the case when grouping geolocations. Instead we rely on a dendrogram, measure heatmap, and interactive graph decomposition of the measure heatmap. These are shown in Figure 2.11, Figure 2.12, and Figure 2.13.

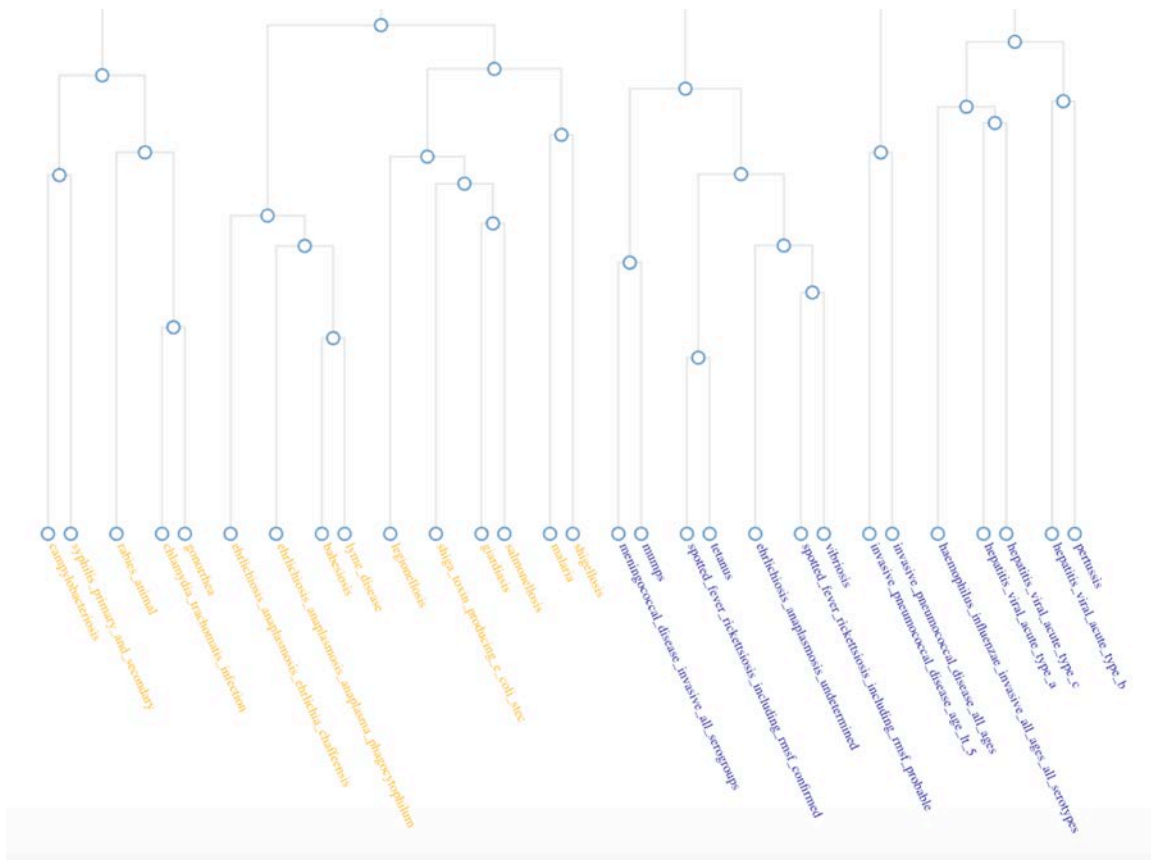


Figure 2.12. Dendrogram of Within-Location Clustering. Tracing diseases from bottom to top can help understand how clusters were merged.

Heatmap Visualization Tool

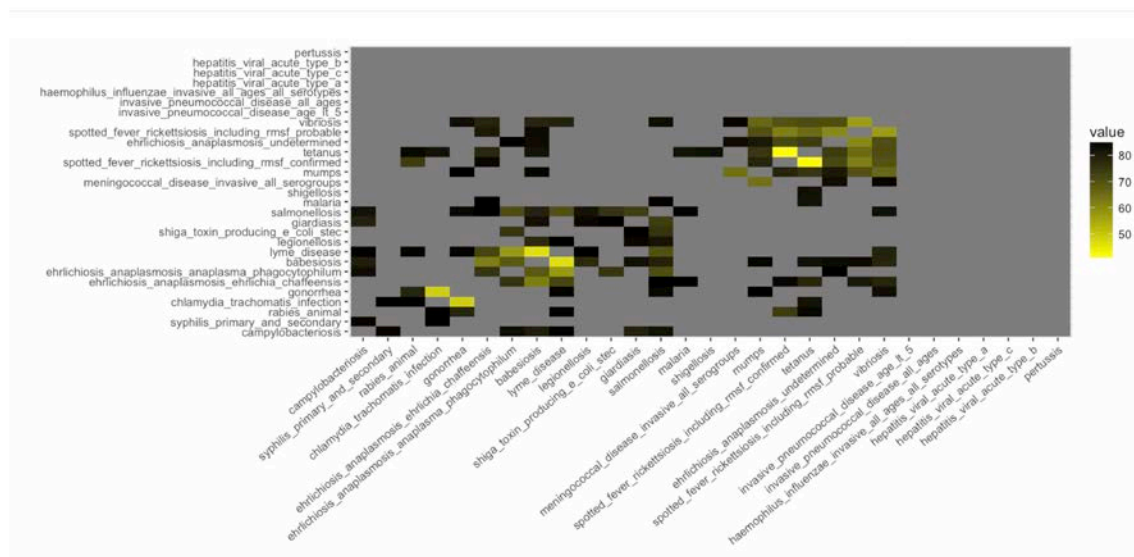


Figure 2.13. Distance Matrix as a Heatmap Visualized within the App. Distances over 86 units have been filtered out (threshold selected interactively within the UI).

The dendrogram (Figure 2.11) identifies cluster assignment as well as providing information for experienced users to assess the number of clusters that might be useful for these data.

The interactive heatmap visualization (Figure 2.12) is typically associated with clustering dendrograms. It allows an experienced user to find pockets of related distances to understand cluster assignment. GTSEDA goes one step further than might be seen in a publication in that the coloring of the heatmap can be thresholded to be removed (colored grey) if it is over a user-selected value.

The value of thresholding is seen in Figure 2.13, which constructs a graph from the heatmap data. GTSEDA places an edge between two diseases if their measurement distance is under a user-defined threshold. The app takes the information in the heatmap and presents it graphically, and because the heatmap value has been incorporated in the presence/absence of an edge, the nodes (diseases or locations) are colored according to color membership. This graphic presentation allows the user to assess the cluster centrality of a node. If a node is only connected to other nodes of the same cluster, then exploratory analysis should be interpreted as that node is well placed and central to the cluster; conversely, if it is on the edge, and connected to an equal number of members from several clusters, then its cluster membership is likely not as well placed. A perfect Ward.D2 clustering scenario would have each cluster having only low distance measures between members of the same cluster, but situation that never happens in practice. This finding is the same interpretation of heatmaps commonly seen in applied clustering literature.

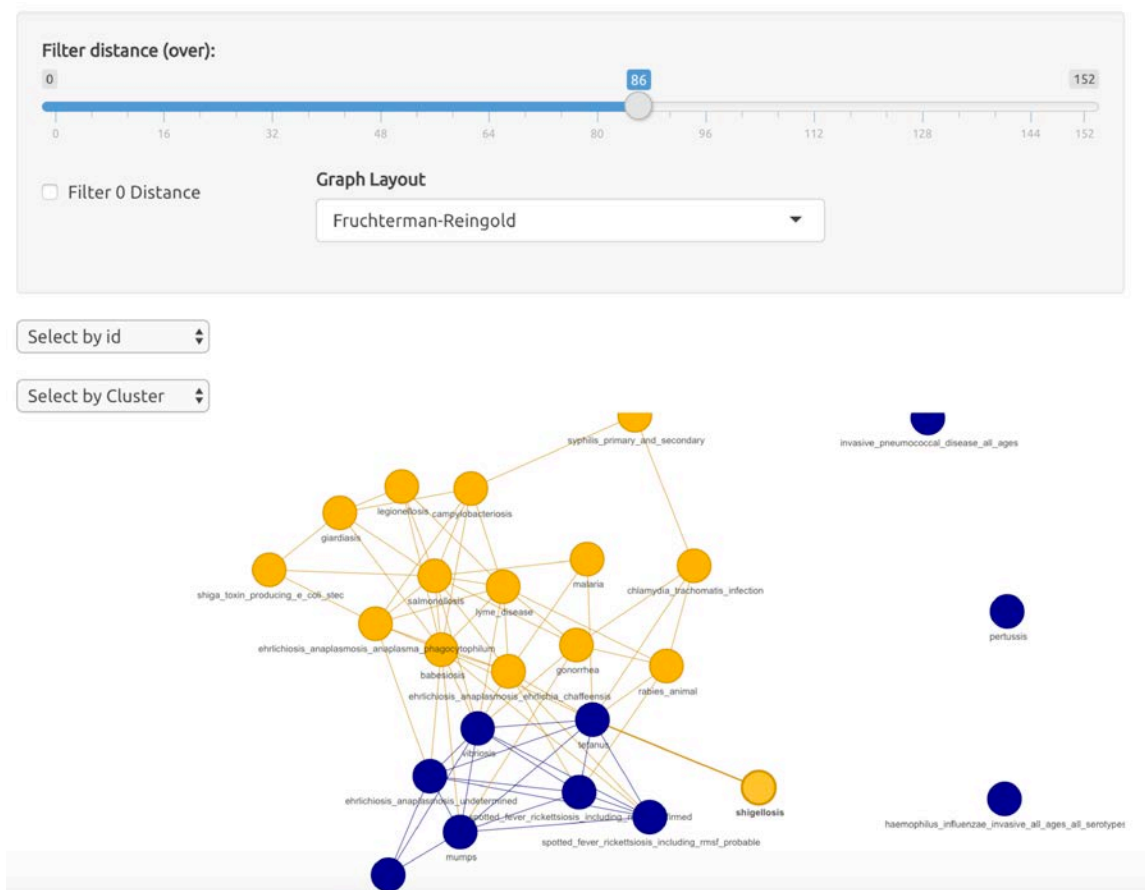


Figure 2.14. The Information of Figure 2.12 Re-interpreted as a Graph. Here edges are allowed between diseases when the heatmap Distance matrix distances are under a user-selected threshold. We

can see that the seasonal bifurcation of New York diseases is preserved in the edges between diseases.

2.2.6 Alarms Tab

Alerting in GTSEDA is done largely through the use of the R package: *surveillance*. The exception to this is Exponentially Weighted Moving Average (EWMA), which is handled internally within the app. The user may select and customize the current alerting system in the Alarms tab (which was formerly called the “Alarming Tab”).

2.2.6.1 Exponentially Weighted Moving Average

EWMA is the default alerting routine within the app. It is not new, and nothing that is discussed here is new. Alerting on EWMA depends on a smoothed historic value exceeding statistically derived upper control lines. The value of smoother is calculated recursively as

$$z_{rt} = \lambda v_{rt} + (1 - \lambda)z_{r(t-1)},$$

so the present value of the smoother is given as $(1 - \lambda)$ times the last value of the smoother plus λ times the current value of the time-series being considered, where $\lambda \in (0,1)$ controls how much emphasis to place on the present observation. If λ is very high, then a strong emphasis is being placed on the current observation. Various methods exist for selecting λ .

Rather than let the user specify the value of λ directly, GTSEDA lets the user input an easier-to-interpret parameterization. We note that that an equivalent mathematical definition is

$$z_{rt} = \lambda v_{rt} + (1 - \lambda)[\lambda v_{r(t-1)} + (1 - \lambda)z_{r(t-2)}]$$

$$z_{rt} = \lambda v_{rt} + \lambda(1 - \lambda)v_{r(t-1)} + \lambda(1 - \lambda)^2 v_{r(t-2)} + \dots + \lambda(1 - \lambda)^N v_{r(t-N)} + \dots,$$

which implies that every observation of the time-series v_r going into the past has a relative weight to the present observation of $(1 - \lambda)^N$ where N is the number of time steps into the past we are considering. We can then define a relative weight and some time period in the past that we want that relative weight to occur. We call this “period in the past” the *span* of the EWMA smoother, and GTSEDA lets the user select a relative weight for that span. If the relative weight is 1/2, this span is typically called a “half-life,” because it is the period of time between halving of the weight.

After a user selects a span, N , and a relative-weight, w , GTSEDA simply assigns $\lambda = 1 - w^{1/N}$, which is a mathematically equivalent parameterization.

For example, the default control options for EWMA in GTSEDA sets a half-life of four weeks. The half-life of four weeks is derived within the app as $\lambda = .159 = 1 - 0.5^{\frac{1}{4}}$. Even though internally EWMA is parameterized in terms of this λ , stating that every four weeks into the past each observation gets half as much weight in the current smoother is a lot more interpretable. Therefore, control options for EWMA include a span and a weight for the smoother.

EWMA Upper Control Line

The upper control line is parameterized by the number of standard deviations outside the mean an EWMA for a time-series should be before generating an alert. Typical values are 3 to 4 standard deviations depending on context.

The upper control line for the EWMA alerting is calculated in a standard way as $UCL_r = \hat{\mu}_r \pm k \hat{\sigma}_r \frac{\lambda}{2-\lambda}$, where $\hat{\mu}_r$ is an estimate of the mean of the current time-series r , $\hat{\sigma}_r$ is an estimate of the standard deviation, and k is a control parameter input by the user. This upper control line follows literature such as that outlined here: <http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm>.

2.2.6.2 Early Aberration Reporting System C1, C2, and C3

After EWMA, the next set of generally applicable alerts for counts are the Early Aberration Reporting System (EARS) algorithms C1, C2, and C3. These algorithms were developed more than a decade ago, and their name refers to names used in the original publications. Fricker et al. (2008) offer a broader discussion and comparison of EARS and cumulative sum (CUSUM)-based methodology. The GTSEDA app uses the implementation of EARS in the R package *surveillance*, which is more broadly discussed in Salmon et al. (2016).

We quote from the *surveillance* R package:

- “In C1 and C2 the expected value is the moving average of counts over the sliding window of the baseline and the prediction interval depends on the standard derivation of the observed counts in this window. They can be considered as Shewhart control charts with a small sample used for calculations.”
- “In C3 the expected value is based on the sum over three timepoints (assessed timepoints and the two previous timepoints) of the discrepancy between observations and predictions, predictions being calculated with the C2 method. This method has similarities with a CUSUM method due to it adding discrepancies between predictions and observations over several timepoints, but is not a CUSUM (sum over 3 timepoints, not accumulation over a whole range), even if it sometimes is presented as such.”

The important control parameter exposed in the UI is “Alpha,” which controls the sensitivity to unusual observations within the time-series before alerting. Alpha is interpreted as the tail probability of observing something as extreme as this under the Null hypothesis of no change in distribution. Lowering Alpha causes alarms to be less frequent and lowers sensitivity.

In addition to EWMA and EARS, the UI has experimental support for other alerting algorithms from the package *surveillance*, but these remaining alerts require a long seasonal time-series.

2.3 Future Work and Discussion

If work on GTSEDA were further supported, then we believe four broad areas of improvement are possible.

2.3.1 Data Sources

In regards to new data source, any data source that is basically syndrome surveillance, or similar in form and schema to geolocated weekly time-series is appropriate to be used within GTSEDA.

Beyond the topic of new data sources, either there must be some constantly re-engineering of GTSEDA to match new data schema or some schema specifications must be standardized within the BSVE such that apps can rely on it in the future.

The BSVE is currently standardizing the concept of a geolocation in the form of a geoJSON project, which was not taken advantage of by GTSEDA, but even beyond that, geolocation standardization is standardizing the storage of time-series for geolocations, such that they can be efficiently queried by all apps.

GTSEDA could bridge the gap by providing a data specification app or UI. Such an app is not GTSEDA *per se* but communicates to GTSEDA via Postgres storage the translation of new data sources to a schema that GTSEDA understands. It would be better to make universal the schema that stores time-series for MMWR weekly data and maps it to geolocations within the BSVE.

Future work in this area is more than just finding new data sources: it is taking part in the standardization of time-series storage within BSVE.

2.3.2 Automation

The lack of automation in the app comes from its origin as “Exploratory Data Analysis” and “Visual Analytics,” but once the topic of alerting is broached, the subject of automation comes close behind.

The systems that state and federal agencies use to monitor syndrome surveillance generally do not ask the user to interact directly with data simply to get an alert. A common paradigm is for the system to asynchronously run a suite of related alerts and annotate MMWR weekly data with the alerts that triggered. For example, in the GTSEDA app, the use case is to specify an alerting algorithm and then return every week to re-run it, but if that is the focus of GTSEDA, then a better idea is to simply run alerting when GTSEDA starts up.

Automation needs to be promoted to a first-class concept within GTSEDA, such that users need to specify very little from week to week but instead only need to specify once what sort of alerts they are interested in receiving.

2.3.3 Modularization

The alerting system within GTSEDA is an example of a module that does not need to be directly placed within the GTSEDA application. There are component parts of the GTSEDA that would fit better within the BSVE if they were decomposed into separate systems that relied on application communications supported and recommended by Digital Infuzion. A direct advantage of this modularization would be that operations that are conceptually distinct would, with some data format standardization, be available to other applications.

Modularization need not be merely breaking GTSEDA into component systems; it can also include re-using available BSVE systems. For example, GTSEDA includes a map widget, and that map widget is partially redundant with the map UI being built into the BSVE. It is redundant in the sense that the display of geoJSON could be accomplished by either UI element. However, GTSEDA also uses the map widget as a primary way to allow the user to navigate from location display to location display, so the GTSEDA needs “onclick” events from the map. This GTSEDA-specific map use is an example of something that becomes more challenging when the obvious components from the BSVE are re-used.

Using more focused expertise in the development of individual modules is another reason to break GTSEDA into component parts. For example, GTSEDA includes a knowledge base for selecting etiologic agents. This knowledge base is a topic unto itself, and BSVE is developing this capability in stand-alone

efforts. Rather than use a purpose-built knowledge base within the app, a more modular approach would query an external knowledge base for etiologic agents.

2.3.4 Further Output Integration

Currently, GTSEDA is on the cusp of having minimal dossier integration. This integration does not include image-based output, which would require an additional one-month effort. In addition, the current organization of the app does not allow users to efficiently communicate the aspects of the output they feel should go to the BSVE dossier. Aspects of the communicating what output is of interest to the user would be easier if the design were broke apart into the modules discussed in modularization.

For example, if the alerting UI were its own module, then it would be clear from the operation of the alerting UI that the context for dossier output was current alerts, and presumably if the map UI is its own module, then it would be clear that the map layer is the dossier output of interest.

We believe that engineering smooth dossier interaction needs to come during modularizing sub-systems.

3.0 Military Biosurveillance: Studying Military Community Health, Well-being, and Discourse through the Social Media Lens

Social media has become a resource for studying different social, emotional, health, and economic conditions of communities through their online activities and shared content. Recently, studies have sought to understand the emotions and behavior in different groups of people through their social media footprints (Blei et al. 2003; De Choudhury et al. 2013). Other studies aim to investigate social issues and phenomena existing in communities through their online activities (Delgado Valdes et al. 2015; Lin 2014).

Social media platforms, such as Twitter, contain publicly available information that provides a resource for potential identification of subpopulations and communities (Blei et al. 2003; De Choudhury et al. 2013; Lin 2014). Applying machine learning and natural language processing techniques to social media content generated by military populations creates a potential to identify, characterize, and monitor their health and well-being. For instance, recent studies used signals from social media to study subpopulations online with the goal of detecting food poisoning within certain subpopulations and geographic regions (Harris et al. 2014) and identifying subpopulations of smokers and drug addicts (Paul and Dredze 2013).

Military service type (e.g., Army, Navy, Marine, Air Force, Active Duty, Reserves, and Veterans) may play a role in the health and well-being of military personnel, including the development of specific health conditions. Boehmer et al. (2003) studied the association between military service and health-related quality of life, using a population-based sample of adults in the United States. They found that the active-duty population had more health complaints than either reserve or veteran populations.

In this work, we aim to understand the differences in online behavior and content produced by military populations, which share common characteristics, such as location, work, and culture, and compare them with non-military populations. Specifically, the goal of this section is to qualitatively and quantitatively estimate language variations and differences in communication behavior across these two populations.

Understanding social media activities and discourse of military populations may help decision makers gain real-time insights into these populations' mental health, including social and emotional stressors, and other health-related issues through a minimally invasive and economic approach. Public health researchers and authorities could use the proposed methods to identify targeted populations quickly and distribute resources effectively.

Next, we list our research questions, provide some background, and describe our data and methods for identifying military users on Twitter. Then, we present our analysis and results. We conclude with a discussion about the implications of our findings.

3.1 Research Questions

Our motivation to study social media activities and discourse of military populations is to better understand their online social interactions and help to identify issues specific to military populations. Overall, we are interested in answering the leading broad questions by addressing the following finer research questions.

- How does the military population use social media?

- RQ 1: What are the differences in tweeting behavior between military and non-military (control) populations?
- What do military users discuss in social media?
 - RQ 2: What are the linguistic differences between the content produced by the military vs. non-military (control) populations?
 - RQ 3: What are the seasonal trends of sentiment expressed in military and control tweets?
 - RQ 4: What kinds of topics do people in the military and non-military (control) populations talk about on social media?
- Do military users talk about health differently than others?
 - RQ 5: Are there any differences in the discourse of health-related topics by the military population compared to the control?

3.2 Background and Related Work

In this section, we first provide a background and summary of prior research about the U.S. military population. Then, we briefly discuss prior work on understanding different populations through social media data.

3.2.1 Characteristics of the U.S. Military Population

The U.S. military consists of active-duty forces (Army, Air Force, Marine Corps, and Navy) and supporting groups (National Guard, Military Reserves, and Coast Guard). Within the United States, armed forces density varies by state; Texas, California, North Carolina, and Virginia have the highest concentrations (Segal and Segal 2004).

In active duty and the reserves, individuals sign up for a specific length of duty and leave service or retire after that term. Three-quarters of the military population are less than 40 years old, and half of the active duty enlisted personnel are less than 25 years old. More than half of the military personnel are married, and 73% of married personnel have children (Office of the Deputy Assistant Secretary of Defense 2013). The military population is diverse, without discrimination of sex, race, or native language. A unique characteristic of the military lifestyle is the frequent relocation of personnel and their families. The military is vulnerable to physical and mental health problems with nearly 18% of active duty deaths caused by illness, and more than 10% of deaths are caused by suicides (Segal and Segal 2004).

3.2.2 Studies on Military Populations

Since the U.S. armed forces changed from drafting to enlistment in 1973, sociologists have debated whether to study the military as an institution or an occupation (Siebold 2001). In general, the military is becoming oriented as a profession yet retains institutional features (Siebold 2001). The U.S. military population reflects America's racial, ethnic, religious, and socioeconomic diversity (Segal and Segal 2004); however, their military status unifies them as a unique subpopulation.

3.2.3 Understanding Populations through Social Media

As more and more users adopt social media, recent studies have attempted to use social media data to understand different subpopulations. Geotagged social media data are being used to identify populations in specific geographical neighborhoods and urban areas (Delgado Valdes et al. 2015; Lin 2014). Another body of work investigates specific demographic groups such as new mothers (De Choudhury et al. 2013), fathers (Blei et al. 2003), and mothers using anonymous social media

platforms (Schoenebeck 2013). These studies use social media profile information to identify users belonging to specific demographic groups or use forums to recruit subjects.

In line with recent research, we seek to study the U.S. military population through the lens of their online social media activities, particularly through Twitter.

3.3 Data

Identifying subpopulations in social media with certain common characteristics (e.g., profession or location) is a challenging task. For our study, the data collection problem entailed differentiating public social media data from the military population and the surrounding civilian population.

Our initial dataset includes nearly 200 million geo-tagged tweets from November 2011 to June 2015 that originated within a 25-mile radius of 31 U.S. military base locations globally. We used this historical dataset to build a lexicon to identify and sample users who are likely to belong to the military population.

For our analysis, we choose six different U.S. military installations located in three states in the continental United States (Table 3.1). We chose locations that have a high ratio of military to surrounding population. For each of these states, we chose one control location that is at least 50 miles away from any military facility and assumed that at this distance users were less likely to belong to the military population. From tweets that originated within a 25-mile radius of military facilities, we sampled users who were likely to belong to the military population using the methodology explained in the next subsection. We sampled the same number of users from non-military locations for our control dataset. In this manner, we collected up to 3,200 of the most recent tweets (in June 2015) per user in the military and control samples. Note that this timeline dataset contains anonymized tweets with and without geographic coordinates.

Table 3.1. Military Locations $L_1 \dots L_6$ and the Corresponding Number of Users Sampled for Both Military and Control Populations Together. The total number of users sampled across six locations is 10,814.

L_1	L_2	L_2	L_4	L_5	L_6
4,246	1,040	1,538	1,372	1,720	926

3.3.1 Data Anonymization

We followed a rigorous data anonymizing procedure to ensure privacy of all Twitter users. The data collected from a social media vendor and through querying the Twitter API were anonymized specifically for usernames, userids, and tweetids. These data were fed into an Elasticsearch engine where they were encrypted using state-of-art encryption algorithms. Our analysis is based only on completely anonymized data, and findings are reported on an abstract, aggregate level. Below is the detailed description of our sampling and data collection procedures.

3.3.2 Sampling Military Users on Twitter

While studying social media activities and content shared by the military population, our first challenge was in sampling Twitter users who are likely to belong to the military. Military population includes individuals who are in active duty, their family members, and veterans. The standard practice in identifying specific events or users in social media is to search for specific terms or hashtags in the tweets [Cui et al. 2012; Starbird et al. 2014]. This approach was not appropriate for our experiments

because we were interested in analyzing the content itself; extracting tweets with such keywords would bias our content analysis.

Another approach often used to identify specific users on social media is to use a database or web listings of users belonging to specific groups (e.g., online listing of Twitter handles of journalists, used in (Soni et al. 2014)). However, to the best of our knowledge, there are no such listings available specifically for military users. Extracting tweet handles for some military organizations from their websites (e.g., @USArmyReserve, @camp_Lejeune, @Military1Source) provided a way to identify only a small subset of military users. Therefore, we devised an approach for discovering potential military user Twitter accounts based on publicly provided content in the profile description.

To gather tweets that have a high likelihood of being posted by someone in the military, we extracted tweets that originated within a 0.5-mile radius from military base locations. These locations were selected based on the highest percentage of military-to-surrounding population ratio obtained from publicly available data¹. The rationale for choosing a 0.5-mile radius was two-fold; it restricted the area and increased the probability of obtaining tweets from military users, and the resulting number of users per area is nearly 1,000, which is a manageable size for faster annotation. Expert annotators classified profile descriptions of these anonymized users and the list of keywords extracted from the classified profile descriptions is shown in Table 3.2.

Table 3.2. Example Keywords Used to Identify Military Users

Group	Keywords
Active Duty	military, national guard, usmc, corporal, sergeant major, hospitalman, sailor, usaf
Family	army wife, usnspouse, military girl, navygirlfriend, army brat, airforce wife
Veteran	veteran, usnveteran, retired army, ex navy

To sample Twitter users who are likely the military population, we extracted tweets from a 25-mile radius of the facilities in chosen military locations and filtered tweets having most of the keywords from our lexicon in their profile description. Because we used both the geo-location and the appearance of keywords in the profile description to sample users, we expect our approach to perform better than the geo-location-based approach used in prior work (Coppersmith et al. 2014). For the control sample, we identified users from the control locations who did not include any of the keywords in their profile description. However, this control sample might include military users if they do not explicitly state their membership in their profile description. Timelines of the sampled users were collected and anonymized according to the description above.

3.4 Analysis and Results

3.4.1 RQ1: Differences in Social Media Activities of Military vs. Control

To identify the differences in tweeting behavior between the military personnel (members and families in the military community) and control, we extracted the following measures: (1) size of their online social networks (i.e., the number of followers and friends), (2) interaction with other Twitter users (using user mentions as a proxy), (3) user's interaction with large groups of virtual communities (using hashtags as a proxy), and (4) ratio of geo-tagged messages to understand their practice of location sharing in social media. We present and contrast the mean counts that represent user activity

¹ <http://www.militaryinstallations.dod.mil/>

and online behavior across populations in Table 3.3. We observe a high degree of variability among the military and control populations.

Table 3.3. Comparing Mean Values for User Activities and Online Behavior across Military vs. Control Populations (p-value ≤ 0.001 ***, p-value ≤ 0.01 **)

Counts	μ_{mil}	μ_{con}	p-value
Favorite	1604.1	2113.8	***
Friend	663.7	498.2	**
Follower	955.9	976.0	
Status	8455.2	8268.4	
Tweet freq.	5.434	6.656	***
Geotag	0.155	0.138	***
Hashtag	0.216	0.186	***
Media	0.097	0.112	***
Mention	0.478	0.497	***
Retweet	0.200	0.239	***
Url	0.237	0.183	***

Twitter Usage and Frequency: We found that tweeting frequency is higher for the control population than the military. The differences in status counts and the number of followers per user are not statistically significant for military vs. control populations. Military users write a higher proportion of tweets with geo-tags. Moreover, it has been reported recently that military personnel are allowed to use smartphones (Powers 2014; BBC 2013), which have the geo-tagging capability.

Size of Social Network and Online Interactions: The mean number of favorite counts is higher for the control population and the mean number of friend counts is higher for the military population.

The mean ratio of tweets with mentions and retweets shows that military users interact less with others on social media using @-mentions and retweets compared to the control, even though they have similar size social networks. However, military users include more hashtags and URLs on average but less media content compared to the control population.

3.4.2 RQ2: Differences in Language Use between Military and Control

Psychology literature suggests that language is a reliable way of measuring people's internal thoughts and emotions (Tausczik and Pennebaker 2010). Hence, we focus on understanding military populations through the language used in their tweets. To identify the differences in the linguistic attributes between the military and control users, we first use a dictionary-based approach applying the psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001).

Social media language, specifically in microblogs, is often found to be non-standard (Eisenstein 2013). Although there are a few recent works on normalizing techniques to convert tweets to more standard language (Han and Baldwin 2011; Yang and Eisenstein 2013), their performance has shown an only marginal increase in accuracy. Because these methods are in a very primary stage of development, we did not perform any normalization. Instead, we used an open vocabulary approach for extracting terms that differentiate the language of the military and control populations in complementary to dictionary-based (LIWC) analysis.

3.4.2.1 Differences in Linguistic Attributes

We used the psycholinguistic lexicon² to measure the differences in linguistic attributes. LIWC consists of several categories of linguistic attributes, such as linguistic or psychological processes, personal concerns, and spoken categories.

To measure the differences in LIWC linguistic categories, we aggregate all tweets per user, then count the number of LIWC terms in each category, and normalize these counts by the total number of tokens in the tweets written by that user. We compare the differences in LIWC terms for the military and control populations using an independent sample t-test. We report the results, showing only the measures that exhibit the same direction in the t-test for all military locations in Table 3.4.

Linguistic Processes: Our results show that the military population uses more *articles* (e.g., a, an, the) and *prepositions* (e.g., to, with, above) compared to the control. Military users talk more about others by using *third-person plural words* (e.g., they, their) in comparison with the control.

Psychological Processes: Military populations talk more about work and financial issues compared to the control populations in social media, as indicated by higher mentions of *work* (e.g., job, majors, labor) and *money* (e.g., bank, income, loan) related terms. In all of the six locations, military users use more *home* (e.g., family, leasing, housing) related words compared to the control, although none of the differences are statistically significant.

Military personnel in certain locations use a significantly higher number of *death*-related terms (e.g., buried, died, kill). Compared to control users, military users talk significantly less about *school*-related terms; they talk less about *religion* (e.g., church, mosque, prayer), although the differences are not statistically significant. Military users in all of the six locations use *inhibition*-related words (e.g., block, constrain, stop) in a significantly higher rate than respective control users.

3.4.2.2 Keyword Extraction

To find keywords that are specific for military and control populations, we extracted terms that differentiate language between these populations. We used a regularized log-odds ratio-based method (Eisenstein et al. 2011), which compares the base word distribution of each group and outputs terms that are specific for each group. We show the top terms for the military and control samples in Table 3.5.

Looking at the top population-specific terms, we find that terms relevant to the events in military life (e.g., Semper [motto of U.S. marine corps], barracks [buildings in military facilities], boot camp, deployed, stationed, Sergeant) are more prevalent in the social media content of the military population. On the other hand, terms related to school, work, and leisure (e.g., ep [episode], tix [tickets], dorm, campus, Raiders [sports], Savemart, Blackstone [stores or businesses], Greensboro, Winston, Sanger [place names]) are more prevalent in the control population's social media content.

²

Linguistic Inquiry and Word Count (LIWC): <http://www.liwc.net>

Table 3.4. Differences in Linguistic Attributes between Military and Control Populations Measured Using LIWC. We only present linguistic categories which have the same directions across populations. $\Delta = (\mu_{mil} - \mu_{con}) \times 10^{-3}$ (p-value ≤ 0.001 ***, p-value ≤ 0.01 **)

	L ₁			L ₂			L ₃			L ₄			L ₅			L ₆		
Category	Δ	t-stat	p	Δ	t-stat	p	Δ	t-stat	p	Δ	t-stat	p	Δ	t-stat	p	Δ	t-stat	p
Linguistic																		
Article	5.4	16.0	***	1.3	1.9		2.2	4.2	**	1.7	3.2		1.6	3.2		5.4	7.4	***
Prepositions	10.0	16.1	***	1.5	1.3		3.3	3.7	*	4.1	4.2	**	3.3	3.7	*	9.4	6.8	***
3 rd -person pl.	0.3	4.0	**	0.5	3.0		0.1	0.7		0.8	5.8	***	0.3	2.62		0.2	1.7	
Psychological																		
<i>Personal</i>																		
Work	1.2	10.9	***	0.3	1.8		0.2	1.7		0.6	2.4		0.5	3		1.3	5.8	***
School	-0.3	-2.8		-2.1	-8.7	***	-1.7	-7.4	***	-1.9	-9.0	***	-1.4	-6.3	***	-0.4	-1.5	
Money	0.8	6.2	***	0.5	1.9		0.1	0.2		0.5	2.2		0.5	2.7		1.4	5.1	***
Home	0	0.3		0.2	1.2		0.5	3.1		0.2	1.4		0	0.2		0	0.1	
Death	0.1	3.7	*	0	0.6		0.2	3.4		0.3	4.7	***	0.1	1.4		0.1	1.2	
Religion	-0.2	-1.2		-1.1	-2.7		-0.2	-0.9		-0.1	-0.6		-0.1	-0.5		-0.3	-0.9	
<i>Relativity</i>																		
Motion	1.1	5.7	***	0.5	1.1		0.4	1.7		1.1	3.7	*	0.6	2.3		0.1	0.4	
Relative	9.2	11.5	***	1.2	0.8		4.7	4.2	**	3.8	3.0		2.3	2		6.6	3.8	*
Space	6.1	14.7	***	1.5	1.8		1.9	3.0		3.0	4.4	**	1.7	2.6		6.5	7.0	***
<i>Cognitive</i>																		
Inhibition	0.7	10.2	***	0.5	3.7	*	0.5	4.7	***	0.7	6.8	***	0.4	4.4	***	0.7	4.7	***
Causation	0.4	2.8		0.6	2.7		0.7	3.7	*	0.8	3.8	*	0.2	0.9		0.6	2.1	
<i>Perceptual</i>																		
Perception	-0.3	-1.5		-0.5	-1.4		-0.4	-1.7		-0.2	-0.5		-0.6	-2.04		-0.3	-0.8	
Spoken																		
Nonfluencies	0.1	3.2		0.1	1.5		0.2	3.5	*	0.1	2.4		0.1	2.2		0	0.5	

Table 3.5. Keywords Specific to Each Military and Control Sample, Extracted Using Sparse Additive Generative Models of Text (SAGE; Eisenstein et al. 2011)

Military	Control
deployed, sergeant, marines, afghanistan, army, soldiers, usmc, stationed, marine, sgt, barracks, #marines, #navy, hooyah, ssgt, pas, oorah, napa, sempa, bragg, #semperfi, bliss, launch, veterans, ty, lejeune, bootcamp, corps, cammies, hooah, airborne, dam, okinawa, deploy, #veterans	pm, dorm, fresno, raider, #wreckem, lbk, tech, lubbock, frfr, burritos, pismo, ttu, ily, como, packs, Ep, shaver, rec, tcu, raiders, que, hp, bojangles, savemar, #texashtech, cheers, stock, campus, shaver

We find that military slang words are widespread in the social media content of military users (e.g., oorah [battle cry of marine corps], hooyah [battle cry of the navy], chow [food]); whereas the control users have widespread usage of internet slang words (e.g., ep [episode], tix [tickets], ik [I know], tbh [to be honest]) and entity names (e.g., Greensboro, Fresno [place names], Bojangles [food chain], ttu [university], Raiders [sports]). These results show that social media language of the military population is different from the control population.

3.4.3 RQ3: Trends of Sentiment for Military and Control

Public opinions about real-world events and concepts may change over time, and opinions are often expressed through social media. Temporal topical analysis and sentiment analysis on social media data are active research areas (Diakopoulos and Shamma 2010; Mei et al. 2007). Additionally, temporal topical analysis is useful for public health research, such as finding disease outbreaks through social media posts (Corley et al. 2010; Culotta 2010). To analyze the seasonal trends of sentiment, we created a temporal dataset by binning tweets from each month. Over a 12-month period, we compared same-user tweet content from military personnel and civilians who wrote at least 10 tweets per month. We used the Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis library (Hutto and Gilbert 2014), a recent rule-based model for sentiment analysis with state-of-the-art performance. For each month, we obtained average scores for positive and negative sentiments of all the users and plotted the overall averages (Figure 3.1).

According to the trend plots in Figure 3.1, the sentiment scores vary across the months of the year for both the military and control populations. Overall, the military population expresses lower positive sentiment in social media compared to the control samples (except for the location L_1). Notably, the positive sentiment scores of military population show an increased trend during the months of November and December, which is the holiday season in the United States (Thanksgiving, Christmas, and New Year).

Looking at the negative sentiment scores in the bottom row of Figure 3.1, the military users express significantly higher amount of negative sentiment in their social media content compared to others. However, the negative sentiment scores show a reverse trend for the two locations L_1 and L_6 , which are located in the same state. Further investigation is needed to understand whether the location of military personnel affects their sentiment expressed in social media.

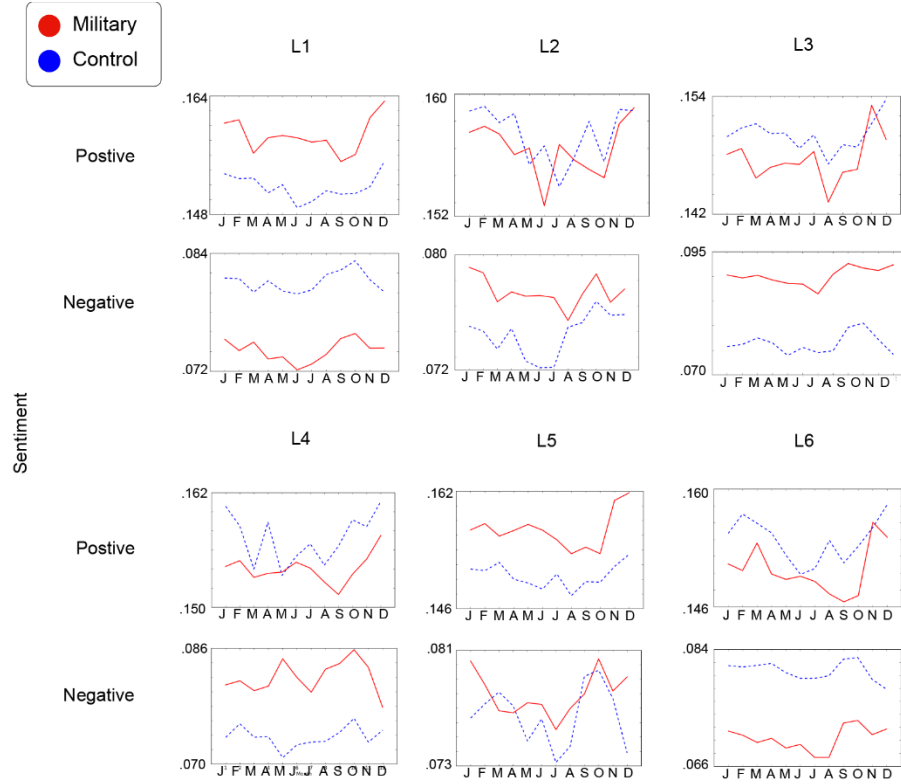


Figure 3.1. Monthly Trend of Positive and Negative Sentiment Scores

3.4.4 RQ4: Topic Variations between Military and Control

Individuals use social media to report about their daily activities, life events, and opinions about various matters. The differences in the topics between the two populations indicate differences in social interactions and broad themes in their daily activities. Therefore, we aim to understand the differences in the language and the latent topics between the two groups.

Latent Dirichlet Allocation (LDA) (Blei et al. 2003) is a classic method for topic modeling. Topic models are based on the assumption that natural language texts are built using a small number of latent topics, and the words in the document represent those topics. LDA is a bag-of-words based generative probabilistic model. The model builds on the words as observed entities, and then it learns the hidden (latent) topics by capturing intra-document statistical structure via mixing distributions of the observed words.

We implemented our topic model using the python library Gensim (Rehurek and Sojka 2010), which is based on online LDA (Blei et al. 2003). After experimenting with several configurations, we determined that 100 topics is a reasonable number of topics. We combined all the tweets from a user and define it as a document unit for topic modeling. We performed standard filtering and cleaning of documents by removing stop words and cleaning HTML tags, followed by lemmatization and stemming. As emojis (smiley faces, sad faces, angry faces) are prevalent in tweets, and they are used to express emotion and other non-verbal cues, we included them in our data.

We trained the topic model using tweets from location L_3 (1,538 documents in the training set, with a 50-50 split between military and control) with 100 topics and used that model to infer topics for the other five military-control pairs. Topics inferred from each document unit are averaged across all the users in

the set to form two distributions for military and control. The averaged distribution across all the topics is compared against the military and control locations.

We selected the topics where control and military populations differ by more than 10% in the weight of their averaged distributions. The relative difference between topic distributions of military and control populations is shown in Figure 3.2. The proportion of the average topic distribution of the military population is shown in the colored area, and the non-colored area represents the respective measure of the control population.

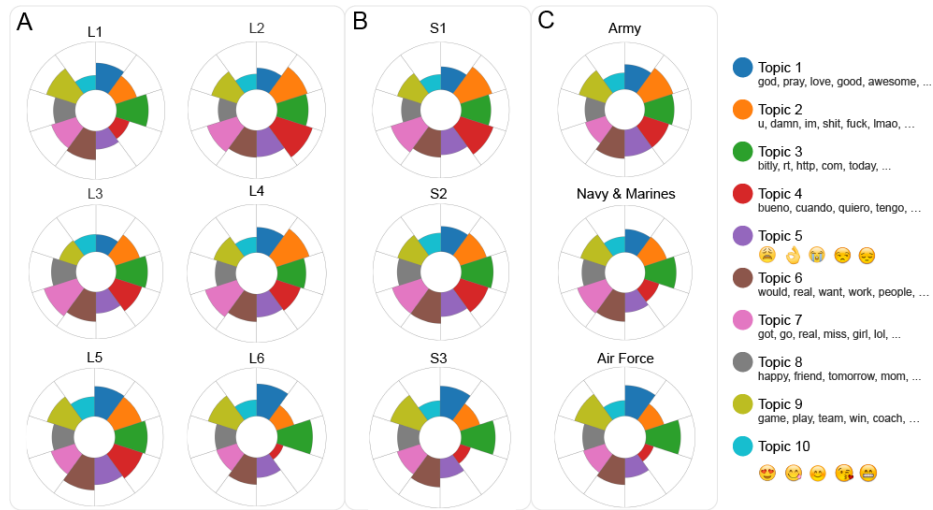


Figure 3.2. A: Distribution of Topics Based on Tweets for Military and Control Populations. B: Distribution of topics based on tweets for military and control populations grouped by geography. C: Distribution of topics based on tweets for military and control populations grouped by military service types. Colored area: Military population; non-colored area: Control population.

Figure 3.2A shows that across all the locations, profanity topic (Topic 2) is more prevalent in the military population compared to the control. Emojis (Topic 5 and 10) are highly prevalent in the tweets from the control population compared to the military users. Other topics do not show any consistent trends across different locations.

To understand the impact of geography and the type of military service (Army, Navy and Marine, and Air Force) on the topics, we look at further groupings. The military populations of states S_1 and S_2 use more Spanish words than the control, but in state S_3 the control uses more Spanish words than the military (Figure 3.2B). We observe the same patterns of increased profanity topics (Topic 2) in the military population and emojis (Topic 5 and 10) in the control population when the populations are grouped by their respective states. When populations are grouped based on military service type, the Navy, Marines, and Air Force use less profanity compared to the control populations, while the Army population uses more profanity compared to their control group (Figure 3.2C).

It is interesting to note that even a completely data-driven model such as LDA can pick up the differences in the social media content of the military and control populations. The differences are present in the topics showing emotional, daily activities, sports, and work-related activities. These findings are consistent with the results observed for our previous research questions.

3.4.5 RQ5: Health-related Discourse between Military and Control

Military populations are considered to be more vulnerable to infectious diseases, such as influenza, and mental health issues, because of overcrowding and a high degree of physical and mental stress (Gray et al. 1999; Pflanz 2001; Russell et al. 2006).

To identify differences in the way military members and their families talk about health conditions compared to the general population, we created a lexicon of health terms and possible misspellings (e.g., “influenza” or “influnza” for the correct spelling “influenza”) and grouped them into six categories as shown in Table 3.6.

Table 3.6. The Example of Health Category Keywords. A * indicates a regular expression; for example fever* indicates words that have a stem fever with difference suffixes such as fevers, feverish and fevered.

Category	Example Keywords and stems
Self-related health experience	suffer*, struggl*, fatigue, weak
ILI-specific symptoms	fever*, cough*, shiver*, runny nose
Disease names and related terms	influenza, sick*, flu, asthma
Health entities	hospital*, doctor*, ER, clinic*
Parts of body & related	lung*, throat, stomach, platelet
Non-ILI-specific symptoms	breath*, diarrhea, dehydrat*, sneez*

We calculated the total counts of terms appearing in user tweets and compared the average term count per token using a t-test (Table 3.7). After Bonferroni correction (Dunn 1961), we find that the direction and significance level of these health measures differ across health-related categories for the military and the civilian populations.

Table 3.7. Comparing the Counts of Health Words for Military vs. Control Populations

Health Category	μ_{mil}	μ_{con}	t-stat	p-value
Self-related health experience	3.04	3.35	-5.246	9.49×10^{-7}
ILI-specific symptoms	71.5	79.0	-3.701	1.29×10^{-3}
Disease names and related terms	1.06	1.17	-4.781	1.06×10^{-5}
Health entities	1.02	1.07	-1.576	6.90×10^{-1}
Parts of body & related	28.5	33.6	-5.134	1.73×10^{-6}
Non-ILI-specific symptoms	49.5	54.5	-3.623	1.75×10^{-3}

Overall, the mean frequency of health-related tweeted complaints from the civilian population is slightly higher than the military population across all health-related categories. These results are obtained through comparisons of military and civilian populations across different geo-locations and military service types. In Table 3.7, we show that civilians use disease-related words more frequently than the military.

3.5 Discussion

In this section, we analyzed social media data collected from military sites and corresponding control populations of users surrounding military locations. We explored the language and metadata inside the tweets from both populations in the following dimensions: behavior, language, and the discourse related to health topics.

Through the analysis of tweeting activities, we found that military populations use fewer retweets and @-mentions compared to the control group. As the usage of retweets and @-mentions are usually considered a measure of social interaction on Twitter (Macskassy 2012), similar to comments and likes on Facebook, these findings indicate that the military users are less interactive on Twitter compared to others.

We found differences in linguistic patterns of military users compared to the control: tweets from military users have a higher usage rate of articles, propositions, third-person plural pronouns, and inhibition words; military users talk more about work and death and less about school-related terms in social media. The increased use of articles suggests that military users use more concrete nouns, and they are interested in objects and things (Tausczik and Pennebaker 2012) compared to the control, while the increased use of propositions suggests that military population is concerned with precision (Tausczik and Pennebaker 2012). Inhibition words are used to suppress strong emotions (Rand et al. 2015). Therefore, increased usage of inhibition words by military users may suggest that they suppress the expression of strong emotional content in social media compared to the control population. Military-specific terms and slang words are prevalent in the tweets of the military users, while the control users talk more about school and leisure activities.

From our analysis, we observed significant differences in online behavior and discourse of the U.S. military when compared to civilian users in social media. Below, we discuss the implications of our findings on life and health of military populations.

3.5.1 Implications for Military Social Life

Our study offers novel and interesting findings on social media activities and the discourse of the U.S. military population. This is an early work towards understanding the role of social networks for improving lives of military populations. From our findings from RQ1, military populations have lower social interactions on Twitter compared to the control users. This finding suggests that military users are socially less active than others in social media. Findings from RQ2 show a significantly higher usage rate of inhibition words, which suggests the self-consciousness expressed in the military populations' messages.

3.5.2 Implications for Military and Public Health

Our findings for RQ3 show significant differences in the use of medically related terms between military vs. civilian populations on Twitter. Overall, civilian populations tend to use more health-related terms (diseases, symptoms, etc.) than military populations. However, we cannot conclude that military populations are healthier compared to civilian populations, as further study is needed to explore the use of colloquial language or military jargon instead of standard disease terminology. Nonetheless, the direction of the variables that indicate health-related terms shows that there are significant differences in the way military and civilian populations talk about health in social media.

The discovered health-related expressions of military personnel on Twitter suggest that it is possible to use social media content from military users to identify emerging health issues that are prevalent in the

military population due to the nature of their job and living conditions. Faster and better identification of health-related issues have implications on public health.

3.5.3 Limitations and Future Work

First, we relied on social media content from Twitter alone to study our research questions. Using only one social media source is a limitation and future work can expand this to other sources such as Facebook and Reddit. Second, we relied on the geotagged tweets for the initial identification of military users. However, recent work shows biases in the geotagged Twitter data regarding text content (Pavalanathan and Eisenstein 2015) and suggests considerations of these biases when generalizing research findings. Third, we relied on geo-origins of the tweets and keywords in Twitter profile descriptions to extract users belonging to the military, but better identification methods can be explored. Fourth, we did not take into account demographics for military and control populations.

There are several directions for future work. Complementing this analysis with an interview study about social media usage of the military users will help researchers and decision makers to understand the limitations in using social media among military personnel.

Moreover, linguistic differences between military and civilian users would enable the construction of classification models to automatically identify military users in social media. Expanding the analysis on health discourse and deriving cues about military health issues to predict disease outbreaks is another possible direction.

Finally, understanding the discourse of military when compared to the civilians helps to identify and prevent social issues affecting them non-evasively. For instance, differences in discourse between military and non-military populations have been effectively used in other studies to identify emotional stress, depression, and post-traumatic stress disorder and related illnesses. In the future, we would like to study fine-grained emotional differences between military and non-military populations over time and model language variations among populations more effectively.

3.6 Conclusion

In this work, we studied language and online behavior of military populations compared to civilians within the same geographic region through social media. We observed significant differences in tweeting behavior between these populations. We further analyzed language inside the tweets and observed that there are significant linguistic differences in emotion and psychological words used between military and civilian populations. Finally, we found that there are significant differences in health-related discourse between the military and civilian populations.

4.0 Military Biosurveillance: Predicting Influenza Dynamics with Neural Networks Using Signals from Social Media

4.1 Motivation

Every year, 500,000 deaths worldwide are attributed to influenza (WHO 2009). The Centers for Disease Control and Prevention (CDC) reports weekly on the level of ILI seen year round in hospitals and doctor visits. These values are used to monitor the spread and impact of influenza; however, by the time the ILI data are released, the information is already 1-2 weeks old and is frequently inaccurate until revisions are made (Paul et al. 2014). To overcome this, we propose making use of large amounts of social media data, such as Twitter, to be a secondary source of information in order to predict current and future ILI proportions—the total number of people seeking medical attention with ILI symptoms. In previous related work, flu forecasting has been accomplished through the use of basic linear autoregressive models, linear autoregression exogenous models, support vector machine regressions, logistic regression classifiers, susceptible, infectious, removed (SIR) models, and more (Broniatowski et al. 2013; Santillana et al. 2015; Riley et al. 2015; Shaman and Karspeck 2012). The addition of social media features to several of these models, such as the linear autoregressive model, has improved the model's performances over ILI data alone (Paul 2016; Smith et al. 2016; Paul et al. 2014). Our work is geared toward applying these data sources to more powerful machine-learning models. Having this predictive power can aid health officials to properly prepare for and respond to yearly flu outbreaks.

4.2 Approach

By integrating the information that people tweet about—e.g., topics, syntax, style and their communication behavior (such as hashtags and mentions)—we built predictive models for ILI and confirmed influenza activity across different geographical locations in the United States. We experiment and evaluate the predictive power of a variety of features and machine-learning models—e.g., support vector machines with radial basis function or linear kernels, AdaBoost with Decision Trees (Pedregosa et al. 2011). We are the first to evaluate the predictive power of neural networks—Long Short-Term Memory (LSTM) for ILI nowcasting and forecasting (Chollet 2015). An LSTM is a special type of recurrent neural network that is capable of preserving information and learning long-term dependencies in data, which traditional recurrent neural networks struggle with. For this specific reason, we chose LSTMs to model our data over the course of several weeks.

4.3 Results

We found that LSTMs achieve the best performance regardless of which text representations are included—e.g., embeddings vs. raw tweets. Of our nine features extracted from Twitter, AdaBoost models learned from unigrams, hashtags, and word embeddings consistently outperform all other features. Using up to four weeks of past data, our models are capable of accurately predicting ILI proportions for the current week and predicting ILI values for up to the next two weeks. We have found that a model tailored to a specific location shows a greater performance than a general model encompassing all regions. In our future work, we will apply our LSTM model to 25 additional locations and combine our ILI and social media data into one predictive LSTM model.

5.0 Chiron Computing Pipeline and Architecture

5.1 Server Architecture and Description

The data and computing architecture for the Chiron BSVE project resided on a PNNL-hosted OpenStack cloud implementation. Software technologies supporting Chiron research included Apache NiFi 0.6.0 and Elasticsearch 1.7.1. NiFi is an Apache Software Foundation project allowing rapid construction of high-performing clusterable Extract Transform Load (ETL) processes. Elasticsearch is a distributed, sharded, Lucene-based NoSQL database, used to allow bulk data storage, search, retrieval, and aggregation.

Three OpenStack machine sizes were used in the construction of the Chiron data and computing architecture (see Table 5.1).

Table 5.1. The Three OpenStack Machine Sizes Used to Build the Chiron Data and Computing Architecture

Type	RAM	VCPU	DISK Space
xLarge	16 GB	8	160 GB
Large	8 GB	4	80 GB
Small	2 GB	1	20 GB

Table 5.2 shows a list of the servers with their types and size of the added volume to allow for extra storage space (if N/A, then no volume is attached):

Table 5.2. Servers Used in the Chiron BSVE Project

Instance (Server) Name	Type	Volume Size
large_httpnode	xLarge	N/A
Nifihost3	xLarge	N/A
data_host	Small	500GB
mongo_host	Large	100 GB
yakutat-1	Large	650 GB
yakutat-2	Large	650 GB
yakutat-3	Large	650 GB
yakutat-4	Large	650 GB
yakutat-5	Large	650 GB
yakutat-6	Large	650 GB
yakutat-7	Large	650 GB
yakutat-8	Large	650 GB
yakutat-9	Large	650 GB
yakutat-10	Large	650 GB
yakutat	Large	N/A

The servers mongo_host and data_host are not a part of the primary data and computing architecture described in this section. Data_host is used as a persistent backup repository for purchased data. Mongo_host houses a MongoDB database containing a subset of the available data; it is used to support occasional one-off research tasks for researchers comfortable in that environment.

Node Nifihost3 hosts the NiFi ETL process.

Servers yakatut-1 through yakatut-10 host Elasticsearch data processes. Server yakatut acts as a cluster head node, where Ansible processes can be executed, and as a single point of interaction for an Apache web server based round robin DNS communicating to individual Elasticsearch data hosts on yakatut-1 through yakatut-10. Large_httpnode is an HTTP only Elasticsearch node, allocated to support memory and compute intensive queries while minimizing potential cluster stability impacts that these queries may have.

There are approximately 270 million social media records captured in the repository, occupying approximately 2TB of disk space across the 10 Yakutat data storage nodes.

All systems are running Centos 7. Ops management tasks are conducted with Ansible, allowing deployment and update of files, services, and users and providing the ability to stop, start, and restart distributed services like Elasticsearch.

5.2 ETL Pipeline

5.2.1 High-level ETL

The ETL pipeline used in the Chiron data and computing architecture is deployed in Apache NiFi 0.6.0, with two custom NiFi processor bundles (NAR format), including nifi-darism-0.4.1 and nifi-elasticsearch-nar-0.4.1. These NARs contain, respectively, message enrichments and a custom processor to allow NiFi to write data to Elasticsearch before version 2.0.

Social media data are delivered through a PNNL commercial contract and deposited on the data_host system. A CRON job periodically copies these files to a landing pad on Nifihost3 and moves the files into an archive filestructure.

Once on Nifihost 3, the ETL pipeline shown in Figure 5.1 is applied.

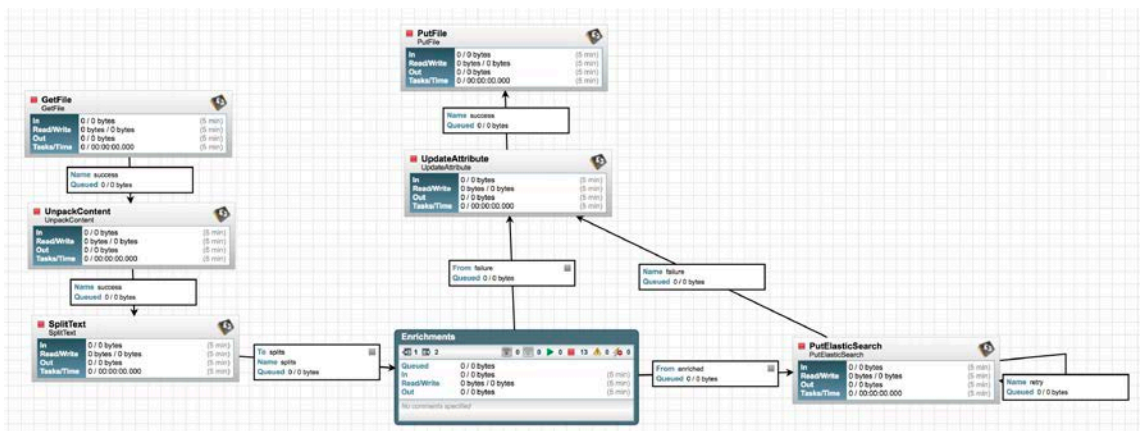


Figure 5.1. The ETL Pipeline

In this pipeline, data are loaded from disk, unarchived, split into individual social media messages, and sent to data enrichment. Messages that fail data enrichment are sent to a failure queue and written to disk for debugging purposes. Successfully enriched messages are written to the Yakatut Elasticsearch cluster

5.2.2 Enrichments

Each message receives the enrichments shown in Figure 5.2.

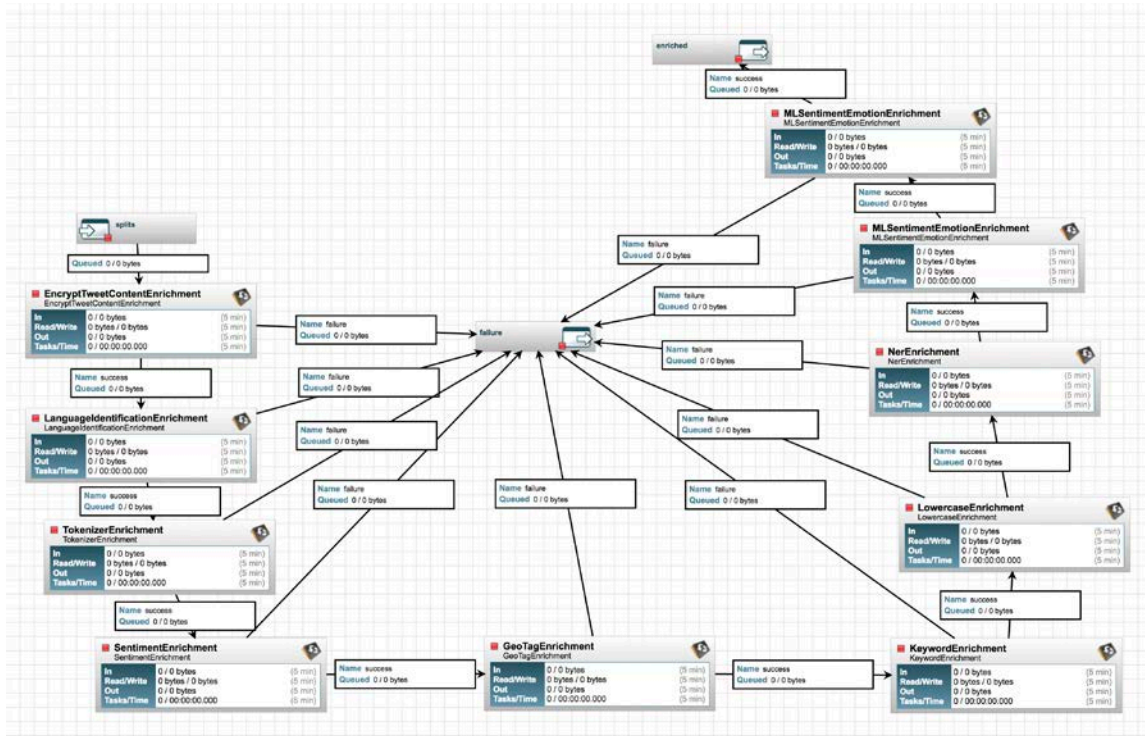


Figure 5.2. Enrichments for the Social Data Analytics Pipeline

In order, these enrichments include:

- **EncryptTweetContent**
 - Encrypts tweet content, including tweet ID, Twitter user IDs, and Twitter usernames, to allow anonymization. Encryption is conducted with the javax.crypto.Cipher library. The AES/ECB/PKCS5Padding algorithm is used. When provided a secret key, identical values will be encrypted to an identical hash, allowing anonymized aggregations of data to be conducted.
- **LanguageIdentificationEnrichment**
 - Uses the carrotsearch implementation of langid.py to do language identification of the tweet body content. References: <https://github.com/carrotsearch/langid-java> and <https://github.com/saffsd/langid.py>.
- **TokenizerEnrichment**
 - Tokenizes sentences and words in multiple languages in preparation for subsequent processors.
- **SentimentEnrichment**

- When a language is identified as Arabic, German, English, Spanish, Persian, French, Italian, Dutch, Dari, Pashto, Russian, Serbian, Turkish, Urdu or Chinese, this processor checks tweet tokens against categorized word lists primarily derived from LIWC lexicons resulting in sentiment and emotion classifications including positive, negative, anger, death, religion, and sadness.
- GeoTagEnrichment
 - Normalizes geocoordinate and does some geo polygon calculations.
- KeywordEnrichment
 - Uses PNNL-developed Rapid Automatic Keyword Extraction algorithm to extract relevant keywords from tweet body.
- LowercaseEnrichment
 - Data storage enrichment, creates lower-case versions of some fields.
- NerEnrichment
 - Uses Stanford Conditional Random Field named entity recognition models to extract the names of people, places, and organizations mentioned in tweets.
- MLSentimentEmotionEnrichment
 - Series of Python-trained machine-learning models exposed in a Java NiFi processor to provide machine-learning-based classification of both emotion and sentiment.

6.0 Understanding Readers' Credibility Perceptions on Social Media Content: Case of Disease Outbreaks on Twitter

Although social media has made information sharing and discovery much easier and faster than before, one of its inherent issues is that not all information is credible and originates from a reliable source. Prior studies have considered just the message text and/or a small subset of potentially influential factors for credibility evaluation. In this section, we present how different reader and author factors affect a reader's credibility perception of information about disease outbreaks on Twitter. Our study results, based on 151 responses, indicate that (1) respondents' credibility perception significantly varies based on the way the information is presented, (2) the influence of an author bio is much more significant than other factors, and (3) a reader's knowledge of topics influences the reliability of credibility assessment.

Our study results show the importance of including available author factors and the reader's domain knowledge when creating human labels for information credibility.

6.1 Introduction

Social media provides a quick and convenient way for individuals and organizations to openly share unlimited packets of information to any and all interested parties. However, one of the inherent issues of an open forum is information credibility, where a single false statement can quickly affect thousands (O'Donovan, et al. 2012). For example, Starbird et al. (2014) reported that many rumors appeared during Boston Marathon Bombing, where a majority of tweets (85–97%) were identified as misinformation. In the context of disease outbreaks, many news articles report the efforts to control the endless spread of misinformation of disease outbreaks in social media; for example, Oyeyemi et al. (2014) showed that 55.5% of tweets related to Ebola were misinformation. Therefore, evaluating social media content to differentiate credible information from misinformation and rumors as well as understanding the characteristics of credible and non-credible information are important.

Fogg and Tseng (1999) described credibility as a perceived quality and believability composed of multiple dimensions. Perceived credibility is subjective and varies depending on the representation of the entity and the characteristics of the person who makes the credibility assessment. As such, much research has attempted to identify important factors and how much each factor would influence a reader's credibility of content in social media. To date, however, prior studies have only considered the message (Ammari and Schoenebeck 2015; O'Donovan 2012; Wang et al. 2015; Xia et al. 2012) or looked at a small subset of potentially influential factors (e.g., user profile image, profile name, location) for credibility evaluation (Morris, et al. 2012; Yang et al. 2013).

In this section, we investigate a reader's credibility perception based on factors pertaining to an author and the reader. Author factors, including author's bio, author's Twitter engagement, attention that author's tweet gained, and characteristics of author's other tweets, were considered. These factors are publicly available, easily, quickly accessed by readers, and may affect their credibility assessment. The reader factors included were age, gender, knowledge of disease outbreaks, and experience in social media. To measure the impacts of these factors, we designed a survey and collected 151 responses.

Our work contributes to a better understanding of the influence of both reader and author factors on credibility perception of the information in social media. Our study results show the importance of including author factors and the reader's domain knowledge when creating human labels for information credibility.

6.2 Related Work

6.2.1 Understanding Credibility in Social Media

A great body of research has investigated information credibility in social media (mostly Twitter) using machine learning. However, the results from prior studies are somewhat limited, because the training data were labeled based only on the text and other influential factors were mostly ignored.

Most machine-learning studies have focused on identifying various features available in social media and obtaining data to build models that classify credible and non-credible information.

Note that, in this paper, a feature refers to a variable used for modeling, whereas a factor points to the one that influences credibility perception. For example, Castillo et al. (2011) used more than 50 features obtained from messages, profiles, topics, and propagations. Similarly, O'Donovan et al. (2012) used 34 features obtained from author's social networks, message content, and behaviors; Xia et al. (2012) used 25 features from author, content, topic, and diffusion in emergency situations; and Wang et al. (2015) used more than 50 features based on profiles, photos, messages, friends, and shares. For this type of research, eliciting unique features and showing greater performance of their classification models (i.e., whether the information is credible or not) is one of the main research goals. However, this way of obtaining human labels on information credibility is somewhat problematic, because only the text messages were presented to human annotators and no other factors were considered. As credibility is a perceived quality and believability comprises multiple dimensions (Fogg and Tseng 1999), humans may not be able to make an accurate and reliable decisions on information credibility from the message only. This is especially true for short tweet messages, consisting of a maximum of 140 characters, making it even more difficult for readers to assess credibility.

The majority of human-computer interaction studies aimed to identify underlying factors that influence one's credibility assessment. For example, Morris et al. (2012) show that users are influenced by heuristics such as message topic, user name, and user profile image when making credibility assessments.

Similarly, Yang et al. (2013) studied the impact of name style, profile image, location, and degree of reader network overlap on credibility perceptions and compared the results among U.S. and Chinese audiences to identify a potential cultural influence. Shariff et al. (2016) presented a correlation analysis of readers' demographics (e.g., gender, age, education, and location) and tweet credibility perception.

However, these research projects only examined the effect of a small subset of all the potentially influential available user factors and/or include reader factors in their study.

In this section, we strive to identify other important factors, from both author and reader, and investigate their impacts on a reader's information credibility perception.

6.3 Study Goals

Based on our literature review findings, our research aims to address the following questions.

- RQ1: What Twitter features influence readers in evaluating their perceived credibility of social media content?
- RQ2: In what ways does a reader's perception of credibility in social media respond to additional contextual information about authors?

- RQ3: What are the impacts of both reader and author factors on a reader’s perceived credibility of social media content?

6.4 Study Design

We designed and conducted a survey study to gain insight into our research questions. The survey consisted of four types of tasks (Figure 6.1). In the first task, we asked participants about their demographics, use of social media, and knowledge of disease outbreaks to capture reader factors (RQ3). In the second task, we asked participants to answer the impact of different Twitter features on their credibility perception (RQ1). In the third task, we asked participants to assess the credibility of 16 different tweets without providing any other information. In the final task, we repeated the third task, but this time, each tweet message was presented with four additional types of information, where each type represented an author factor that we were interested in (RQ2,3).

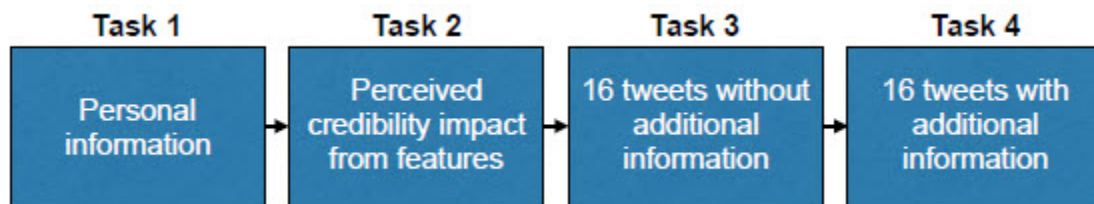


Figure 6.1. Procedure of the Survey. After answering personal information, respondents were asked to complete four types of tasks. The same 16 tweets were used in Tasks 3 and 4, and the corresponding 16 tweet pages were randomly assigned to survey respondents.

Regarding the fourth task, we designed the tasks to capture the following four author factors: (1) Attention: How much attention does the author’s tweet receive? (2) Self-description: Does the author’s biographical text describe him/herself? (3) Engagement: How much does the author engage in Twitter? (4) Professionalism: Are the author’s other tweets personal or professionally written tweets? Again, these are the factors that are publicly available, easily, quickly accessed by readers, which we believe may affect the readers’ credibility assessment.

Figure 6.2 illustrates an example of the tweet page presented with four factors. Each author factor has two conditions, making sixteen tasks (2x2x2x2) in total. Sixteen different, real tweets about Zika, Ebola, and Chikungunya were used for each task (blue box in Figure 6.2). To focus the participants’ credibility judgments on just the four factors, two domain experts reviewed and confirmed that the tweets looked professionally written and described plausible information, which could be true or false. For all questions, including perceived credibility of the tweet, we applied a 5-point Likert scale (1 – not credible; 5 – very credible).

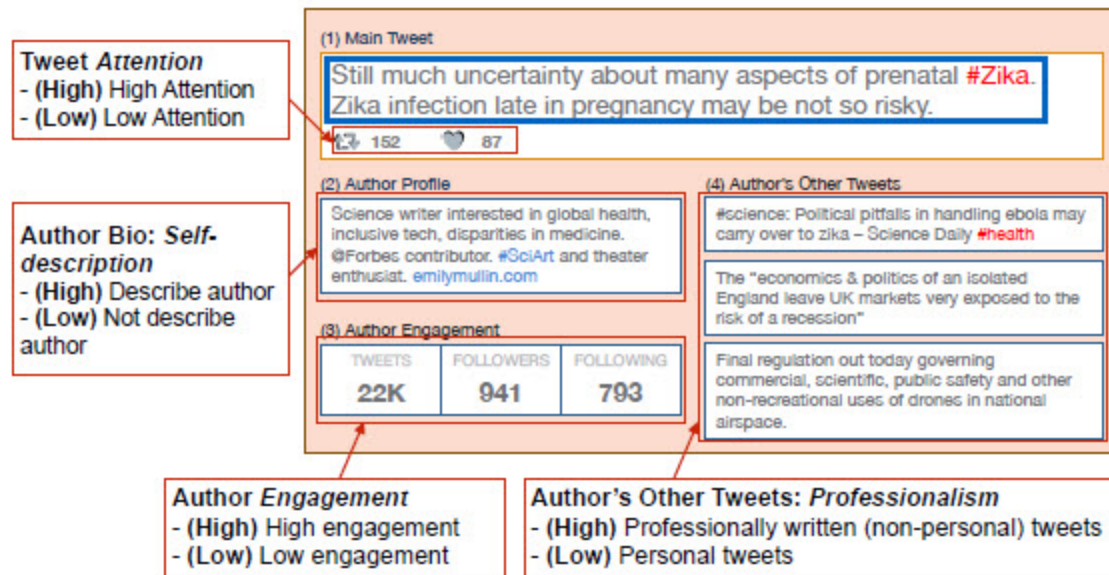


Figure 6.2. Example of the Tweet Page Used in Task 4. Four author factors (two conditions per factor) were measured. We have 200,000 samples of users who posted tweets related to disease outbreaks, and High and Low conditions in Tweet Attention and Author Engagement were decided based on the medians of each case from our samples. Note the High and Low conditions for each factors.

After obtaining an approval from our Internal Review Board, we distributed the survey to people via mailing lists of our organization, word of mouth, and social media/networks. The survey took approximately 15-20 minutes to complete and we collected 162 responses. We excluded 11 incomplete responses and analyzed the remaining 151 survey responses.

6.5 Results

Our data analysis was based on 2,416 individual questions for Task 3 and Task 4 each. We fit a mixed-effects analysis of variance model with a normal conditional distribution and random effects for repeated measures to account for the non-independent nature of the data.

6.5.1 Survey Respondent Demographics

Our survey respondents consisted of 73 males and 78 females. Sixty-two respondents were between 20-30 years old, 51 were between 30-40 years old, and 38 were more than 40 years old. The average number of years someone participated in social media—e.g., Twitter—was 2-3 years. Respondent knowledge of disease outbreaks was fairly high, averaging 3.78 (3: neutral, 4: somewhat familiar) out of 5.

6.5.2 RQ1: Impact of Features on Perceived Credibility

We measured the impact of nine author-based (Figure 6.3) and ten tweet-based (Figure 6.4) features on a reader's credibility perception. For author features, the author's bio showed the greatest impact (Mean: 4.22) followed by having primary topics from author's tweets (3.92). This indicates that the respondents considered what and how the author describes him/herself in the bio and what the author says on Twitter

to be very important. Author's profile image and name, which were identified as main factors in prior studies (Morris et al. 2012; Yang et al. 2013), did not show much impact in our study.

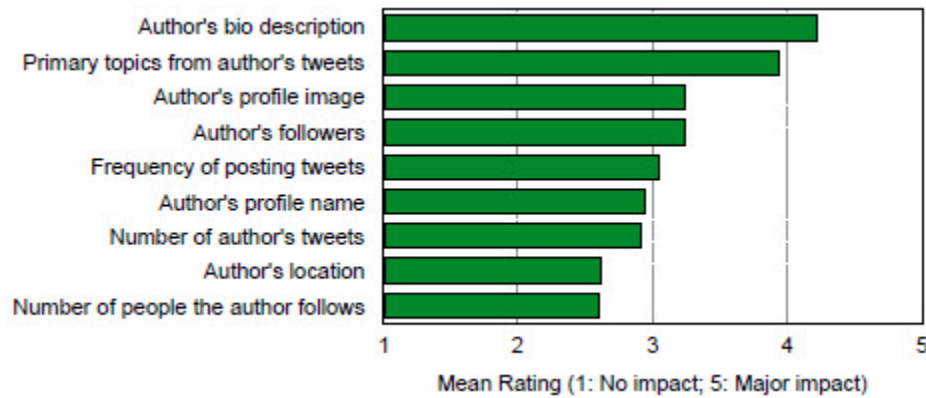


Figure 6.3. Impacts of Author-based Features on Credibility Assessment

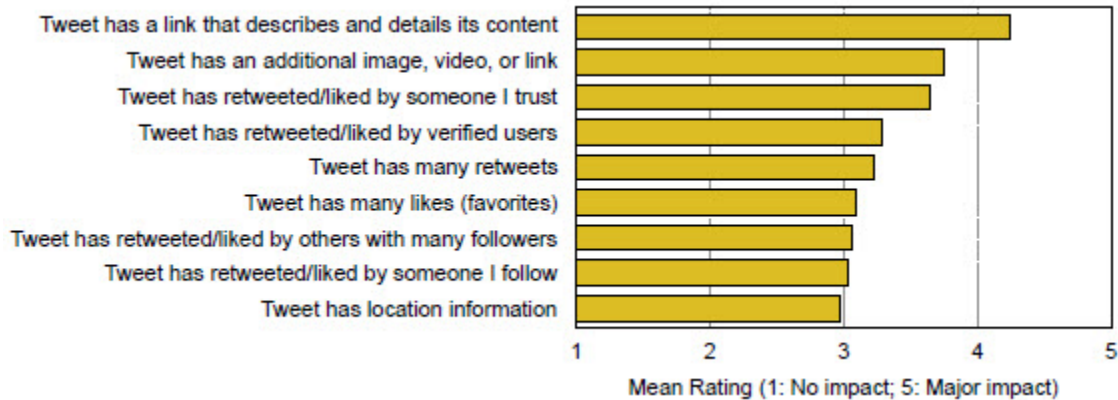


Figure 6.4. Impacts of Tweet-based Features on Credibility Assessment

Regarding the tweet features, tweets containing a link that details tweet content was ranked the highest (4.19). The second highest feature was having an additional image, video, or link (3.77), similar to the highest-ranked feature. Because of limited space in a single tweet, the respondents may have felt that additional information sources increase the credibility of the tweet. The results also indicated that other users who retweeted or liked the tweet are the important factors on credibility. An interesting insight here relates to the types of “other users,” as there is a difference among users they trust (3.65), users verified by Twitter (3.21), and users they follow (3.09). We see that the respondents rated users they trust higher than verified users, emphasizing a subjective nature of credibility perception.

6.5.3 RQ2: Variance in Perceived Credibility

When comparing the results of credibility assessments between Task 3 (tweets only) and Task 4 (tweets with other info), we identified differences for all 16 tweets (Figure 6.5). Additional information caused a significant difference ($p < 0.05$) in perceived credibility in 13 tweets (81%), where 9 of them decreased in Task 4. This result indicates that one's credibility perception can be highly influenced by other factors, stressing the importance of considering various factors and giving the readers more contextual information when making credibility assessments.

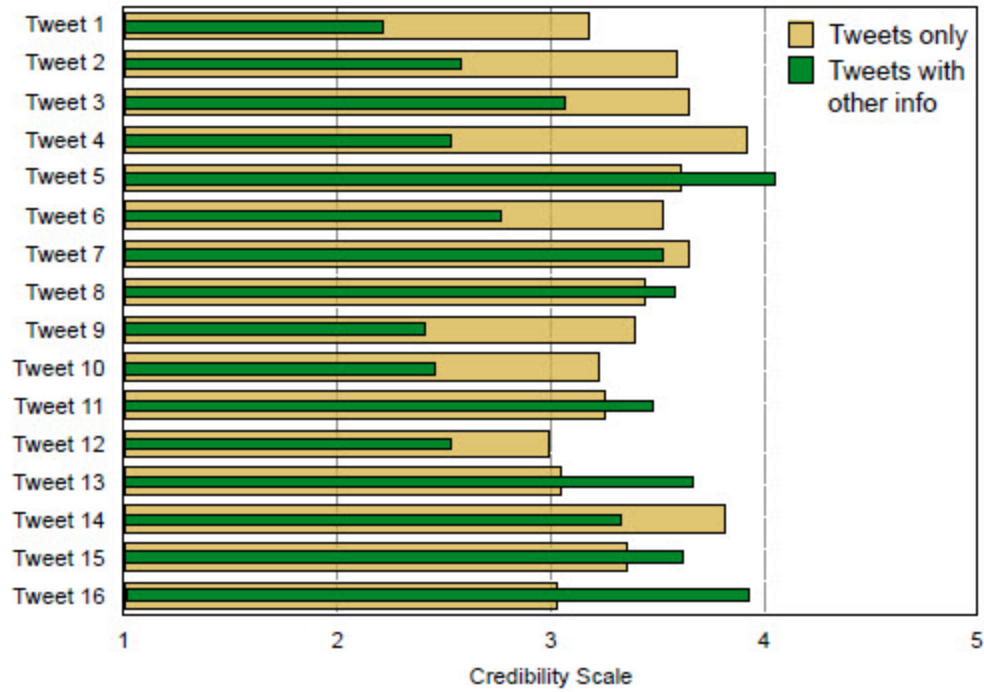


Figure 6.5. Difference in Credibility Assessments between Task 2 and Task 3 for the Same 16 Tweets. Overall, respondents' credibility judgments were significantly influenced by other factors ($p < 0.05$).

6.5.4 RQ3: Reader and Author Factors on Perceived Credibility

We found significant impacts of specific author and reader factors on a reader's credibility perception (see Table 6.1). For author factors, all four factors showed significant influence, yet the impact of the author's bio was much higher than any other factors. This finding implies that the level of detail an author uses to describe him/herself is a critical factor in readers' credibility perception. Figure 6.6 shows the difference between two conditions for each factor. Note that Low in the author bio means there is no description about the author, and Low in the other tweets means there are only personal (non-professionally written) tweets.

Table 6.1. Summary of the Influence of Author and Reader Factors on Credibility Perception. All author factors showed significant influences, but the impact of author bio was incomparable to other factors.

Type	Factor	F-value	Df	Sig (p-value)
Author	Author Bio	421.83	1	0.000
	Other Tweets	67.72	1	0.000
	Author Engagement	25.71	1	0.000
	Tweet Attention	7.37	1	0.007
Reader	Topic Familiarity	4.07	2	0.017
	Age	1.58	2	0.206
	Gender	0.97	1	0.323
	Twitter Experience	0.21	4	0.933

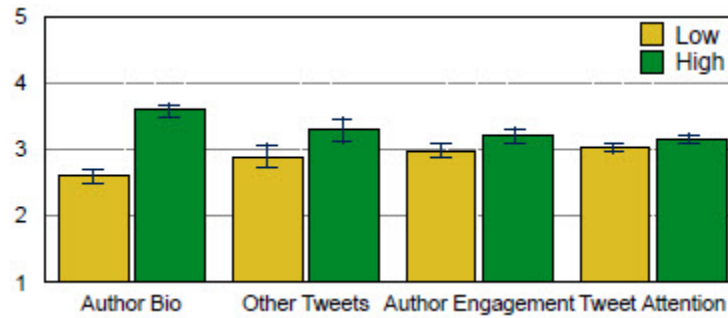


Figure 6.6. Difference in Credibility Ratings in Task 4 between Two Conditions for Each Author Factor. All factors showed significant differences ($p < 0.05$). Note that Low in the author bio refers to no author description, and Low in the other tweets refers to non-professional tweets.

We further measured the impact of each author factor on the changes of credibility perception with respect to the difference between Task 3 and 4 as a dependent variable (Figure 6.7). In general, the respondents gave lower rates to the same tweets in Task 4. The author description exhibited the most significant impact on it, and we found only one case where the rating increased when there is an author description in the bio.

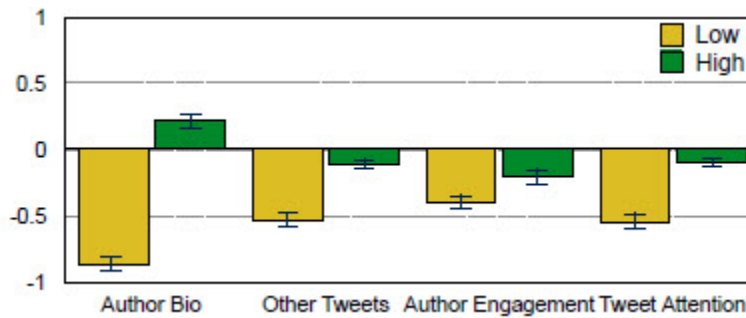


Figure 6.7. Difference in Credibility Ratings between Task 3 and 4 (all $p < .05$). Note that Low in the author bio refers to no author description, and Low in the other tweets refers to non-professional tweets.

This result again indicates the importance of author information in credibility assessments. For reader factors, only topic familiarity showed a significant impact (Table 6.1). Respondents with high domain familiarity (Mean: 3.47 in Task 3; 3.14 in Task 4) exhibited higher credibility perception than the low group (3.28 in Task 3; 2.97 in Task 4; $p < .05$). We also found the same result from the responses in Task 3, which indicates a strong influence of topic familiarity on credibility assessments.

Lastly, we extended the comparison analysis of credibility assessments among topic familiarity groups by looking into reliability (i.e., consistency) of their assessments. To measure this, we used the standard deviations of the mean scores for each group. For the responses in Task 3, the high topic familiarity group was smaller (1.20) than the middle (1.30) and low (1.22) groups. The responses in Task 4 showed greater differences: high (1.15), middle (1.24), and low (1.23). This result indicates that there is generally less variation in the credibility assessment from people who are more familiar with the topic of the tweets (high) and potentially more variability in people who know only a small amount (middle).

6.6 Discussion and Conclusion

This study has shown evidence that various Twitter features influence a reader's information credibility assessments. We presented the influence of both author and reader factors on a reader's credibility perception on social media content related to disease outbreaks. We first showed the significant differences between Task 3 and Task 4, indicating a reader's credibility perception can be highly influenced by other factors. This finding emphasizes a careful design of measuring information credibility, which has not been well considered in many prior studies. We specifically focused the influence of four author factors (two conditions each) and four reader factors on a reader's credibility perception. Although all four author factors showed significant influences, the author bio exhibited significant influence compared to other factors and was the only factor that increased credibility assessment in Task 4 (Figure 6.7). This is consistent with what the respondents indicated in Task 2. Because all four factors were important, the result still emphasizes the importance of making various author's credentials accessible at a glance to accurately measure content credibility.

In fact, some of the features, such as a bio length, bio word count, number of tweets, favorites, followers, and followings, which pertain to the four author factors, have been used in classification modeling (Castillo et al. 2011; O'Donovan et al. 2012; Wang et al. 2015). In addition, our study results suggest using corresponding additional features derived from the author's bio (e.g., does the author's bio describe her interest, job, etc.?) and other tweets (e.g., are the author's other tweets consistent with respect to readability or topic?, what is the proportion of professionally written tweets?) to build models for classifying credible and non-credible information.

Finally, we found that topic familiarity from reader factors influenced the reader's credibility perception. Respondents with high topic familiarity exhibited more positive credibility perception than those who were not familiar with the topic. More importantly, those high familiarity group respondents were more consistent in credibility assessment. The reliability of assessment is important, because not every tweet has ground truth, which sometimes could be difficult even for the high familiarity group respondents to evaluate social media content correctly. Especially for the information related to disease outbreaks, collecting and using credibility evaluations that yield a high degree of consensus from multiple evaluators will be important. While our study offers a number of insights, the results may not be generalized, because we only considered tweets related to disease outbreaks. Prior studies indicate that people's credibility judgment can vary depending on topics (Morris et al. 2012; Shariff et al. 2016; Yang et al. 2013); thus, our study results might be influenced by the topic of disease outbreaks. Therefore, our future work will be to study if our findings are reproducible across different topics. We also would like to apply our insights into classification models and evaluate their performance on credibility assessment. In summary, our work contributes to a better understanding of the influence of different author and reader factors on a reader's credibility perception and to providing insights on extracting new features and using them in machine-learning training models for classifying credible and non-credible information.

7.0 Software Delivered

- Generalized Time Series Exploratory Analysis app is now live and available in the BSVE app store.
- SODA POP time series app
 - http://chiron-shiny.pnnl.gov/shiny/roun308/soda_pop/
 - user: shiny_app
 - password : shiny_app_1234!
- Military facility ILI app
 - http://chiron-shiny.pnnl.gov/shiny/roun308/mil_fac_ili/
 - user: shiny_app
 - password : shiny_app_1234!
- DARISM Social Media Analytics Pipeline

8.0 Publications

[in PA review] Volkova, S, E Ayton, K Porterfield, and CD Corley. “Forecasting Influenza-like Illness Dynamics for Military Populations using Neural Networks and Social Media” in prep for PLoS ONE.

[In Review] Anderson, Aryk, K Shaffer, A Yankov, CD Corley, and NO Hodas. “Beyond Fine Tuning: A Modular Approach to Learning on Small Data”. Submitted to ICLR 2017.

[Accepted] Ayton, Ellyn and Volkova, S. “Predicting Influenza Dynamics with Neural Networks Using Signals from Social Media” in Women in Machine Learning 2016.

[Accepted] Rounds, J and CD Corley “Soda Pop: A Time-Series Clustering, Alarming and Forecasting App in the Biosurveillance Ecosystem” ISDS 2016

[Accepted with revisions] Volkova, S, LE Charles-Smith, J Harrison, and CD Corley. “Uncovering the Relationships Between Military Community Health and Affects” submitted to EJP Data Science

[Accepted] Pavalanathan U, Datla VV, Volkova S, Charles-Smith LE, Pirrung MA, Harrison JJ, Chappell AR, Corley CD 2016. “Studying Military Community Health, Well-being, and Discourse through the Social Media Lens.” *Lecture Notes in Computer Science*

Poster and flash talk for BSVE TIM 2016 meeting.

[Accepted] Pavalanathan, U, V Datla, S Volkova, LE Charles-Smith, J Harrison, M Pirrung, A Chappell, CD Corley. “Discourse, Health and Well-being of Military Populations through the Social Media Lens” AAAI W3PHI, Phoenix, AZ, Feb 2016.

[Accepted, Lightning Talk] Charles-Smith LE, AG Rittel, U Pavalanathan, and CD Corley. “Towards Influenza Surveillance in Military Populations using Novel and Traditional Sources.” Abstract in the International Society for Disease Surveillance, Denver, CO. December 2015.

[Accepted] W Smith, AR Chappell, and CD Corley. “Medical and Transmission Vector Vocabulary Alignment with Schema.org” Proceedings of the International Conference on Biomedical Ontology. July 2015

9.0 References

- Ammari, T and S Schoenebeck. 2015. "Understanding and Supporting Fathers and Fatherhood on Social Media Sites." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. pp. 1905-1914. ACM.
- BBC. 2013. "US Military Approves Android Phones for Soldiers." British Broadcasting Corporation. <http://www.bbc.com/news/technology-22395602>
- Blei DM, AY Ng, and MI Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3, 993-1022.
- Boehmer, TK, VL Boothe, WD Flanders, and DH Barrett. 2003. "Health-related Quality of Life of U.S. Military Personnel: A Population-based Study." *Military Medicine* 168(11), 941-7, <http://search.proquest.com/docview/217055081?pq-origsite=gscholar>
- Broniatowski, DA, MJ Paul, and M Dredze. 2013. "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic." *PloS ONE* 8.12 (2013): e83672.
- Castillo, C., M. Mendoza, and B. Poblete. 2011. "Information Credibility on Twitter." In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 675–684.
- CDC. 2009. "Assessment of ESSENCE Performance for Influenza-like Illness Surveillance after an Influenza Outbreak—U.S. Air Force Academy, Colorado." 2009. Centers for Disease Control, *MMWR Morb Mortal Wkly Rep*, 60(13): 406-9.
- Chollet, F. *Keras*. 2015. github.com/fchollet/keras.
- Coppersmith, G, C Harman, and M Dredze. 2014. "Measuring Post Traumatic Stress Disorder in Twitter." In *Proceedings of ICWSM*, May 2014. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8079>
- Corley, CD, DJ Cook, AR Mikler, and KP Singh. 2010. "Text and Structural Data Mining of Influenza Mentions in Web and Social Media." *International Journal of Environmental Research and Public Health* 7(2), 596-615
- Cui, A, M Zhang, Y Liu, S Ma, and K Zhang. 2012. "Discover Breaking Events with Popular Hashtags in Twitter." In *Proceedings of CIKM*, pp. 1794-1798. ACM, New York, NY, USA. <http://doi.acm.org/10.1145/2396761.2398519>
- Culotta, A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. 2010. In *Proceedings of the First Workshop on Social Media Analytics*, pp. 115-122. ACM.
- De Choudhury, M, S Counts, and E Horvitz. 2013. "Predicting Postpartum Changes in Emotion and Behavior via Social Media." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3267-3276. ACM.
- Delgado Valdes, JM, J Eisenstein, and M De Choudhury. "Psychological Effects of Urban Crime Gleaned from Social Media." In *Proceedings of ICWSM* (Apr 2015). <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10563>

- Diakopoulos, NA and DA Shamma. 2010. "Characterizing Debate Performance via Aggregated Twitter Sentiment." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1195-1198. ACM.
- Dunn, OJ. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56(293), 52-64.
- Eisenstein, J, A Ahmed, and EP Xing. 2011. "Sparse Additive Generative Models of Text." In *Proceedings of ICML*, pp. 1041-1048. Seattle, WA. http://www.icml-2011.org/papers/534_icmlpaper.pdf
- Eisenstein, J. 2013. "What to Do about Bad Language on the Internet." In *Proceedings of NAACL*, pp.359-369. ACL, Stroudsburg, Pennsylvania. <http://www.aclweb.org/anthology/N13-1037>
- Fairchild GC. 2014. "Improving Disease Surveillance: Sentinel Surveillance Network Design and Novel Uses of Wikipedia." In *Iowa Research Online*, University of Iowa.
- Fogg, BJ and H. Tseng. 1999. "The Elements of Computer Credibility." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 80–87.
- Fricker, RD, BL Hegler, and DA Dunfee. 2008. "Comparing Syndromic Surveillance Detection Methods: EARS versus a CUSUM-based Methodology." *Statistics in Medicine* 27:3407-3429.
- Gray, GC, JD Callahan, AW Hawksworth, CA Fisher, and JC Gaydos. 1999. "Respiratory Diseases among US Military Personnel: Countering Emerging Threats." *Emerging Infectious Diseases* 5(3), 379.
- Han, B and T Baldwin. Lexical Normalisation of Short Text Messages: Makn sens a# twitter. In *Proceedings of HLT-Volume 1*, pp. 368-378. ACL.
- Harris, J, R Mansour, B Choucair, J Olson, C Nissen, J Bhatt, and S Brown. 2014. "Health Department Use of Social Media to Identify Foodborne Illness—Chicago, Illinois, 2013-2014." *Morbidity and Mortality Weekly Report* 63(32): 681.
- Hutto, CJ and E Gilbert. E. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In Eighth International AAAI Conference on Weblogs and Social Media. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109>
- Lin, YR. 2014. "Assessing Sentiment Segregation in Urban Communities." In *Proceedings of the 2014 International Conference on Social Computing*, pp. 9:1-9:8. ACM, New York, NY, USA (2014). <http://doi.acm.org/10.1145/2639968.2640066>
- Macskassy, SA. 2012. "On the Study of Social Interactions in Twitter." In *Proceedings of ICWSM*.
- Mei, Q, X Ling, M Wondra, H Su, and C Zhai. 2007. "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs." In *Proceedings of the 16th International Conference on World Wide Web*, pp. 171-180. ACM.
- Morris, MR, S Counts, A Roseway, A Hoff, and J Schwarz. 2012. "Tweeting is Believing?: Understanding Microblog Credibility Perceptions." In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 441–450. ACM.

O'Donovan, J, B Kang, G Meyer, T Hollerer, and S Adalii. 2012. "Credibility in Context: An Analysis of Feature Distributions in Twitter." In *Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom)*, pp. 293–301. IEEE.

Office of the Deputy Assistant Secretary of Defense. 2013. *Demographics Report*.
<http://download.militaryonesource.mil/12038/MOS/Reports/2013-Demographics-Report.pdf>

Oyeyemi, SO, E Gabarron, and R Wynn. 2014. "Ebola, Twitter, and Misinformation: A Dangerous Combination?" *British Medical Journal*.

Paul, MJ and M Dredze. 2013. "Drug Extraction from the Web: Summarizing Drug Experiences with Multi-dimensional Topic Models." In *Proceedings of HLT-NAACL*, pp. 168-178.

Paul, MJ, A Sarker, JS Brownstein, A Nikfarjam, M Scotch, KL Smith, and GA Gonzalez. 2016. "Social Media Mining for Public Health Monitoring and Surveillance." In *Proceedings of Pacific Symposium on Biocomputing*, Vol. 21.

Paul, MJ, M Dredze, and D Broniatowski. 2014. "Twitter Improves Influenza Forecasting. *PLOS Currents Outbreaks*.

Pavalanathan, U and J Eisenstein. 2015. "Confounds and Consequences in Geotagged Twitter Data." In *Proceedings of EMNLP*, pp. 2138-2148. ACL, Lisbon, Portugal, September 2015.
<http://aclweb.org/anthology/D15-1256>

Pedregosa F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, and D Cournapeau. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12, pp. 2825-2830.
<http://scikit-learn.org/stable/index.html>

Pennebaker, JW, ME Francis, and RJ Booth. 2001. "Linguistic Inquiry and Word Count." *LIWC 2001*. Mahway: Lawrence Erlbaum Associates 71.

Pflanz, S. 2001. "Occupational Stress and Psychiatric Illness in the Military: Investigation of the Relationship between Occupational Stress and Mental Illness among Military Mental Health Patients." *Military Medicine* 166(6): 457.

Powers, R. 2014. "Can I Use My Cell Phone During Basic Training?" *The Balance*.
<http://usmilitary.about.com/od/armyjoin/a/basiccellphone.htm>

Rand, DG, G Kraft-Todd, and J Gruber. 2015. "The Collective Benefits of Feeling Good and Letting Go: Positive Emotion and (dis)Inhibition Interact to Predict Cooperative Behavior." *PloS ONE* 10(1), e0117426.

Rehurek, R and P Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of LREC 2010 Workshop: New Challenges for NLP Frameworks*, pp. 46-50. University of Malta, Valletta, Malta. <http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf>

Riley, P, M Ben-Nun, JA Linker, AA Cost, JL Sanchez, D George, DP Bacon, and S Riley. 2015. "Early Characterization of the Severity and Transmissibility of Pandemic Influenza Using Clinical Episode Data from Multiple Populations." *PLoS Comput Biol* 11.9: e1004392.

- Russell, KL, MP Broderick, SE Franklin, LB Blyn, NE, Freed, E Moradi, DJ Ecker, PE Kammerer, MA Osuna, AE Kajon, CB Morn, and MAK Ryan. 2006. "Transmission Dynamics and Prospective Environmental Sampling of Adenovirus in a Military Recruit Setting." *Journal of Infectious Diseases* 194(7): 877-885.
- Salmon, M., Schumacher, D. and Höhle, M. 2016. "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance." *Journal of Statistical Software* 70(10): 1-35.
- Santillana, Mauricio, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, and Brownstein JS. "Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance." *PLoS Comput Biol* 11.10: e1004513.
- Schoenebeck, SY. 2013. "The Secret Life of Online Moms: Anonymity and Disinhibition on YouBeMom.com." In *Proceedings of ICWSM*.
- Segal, DR and MW Segal. 2004. "America's Military Population." *Population Bulletin*, vol. 59. Population Reference Bureau, Washington, DC.
- Shaman, J and A Karspeck. 2012. "Forecasting Seasonal Outbreaks of Influenza." In *Proceedings of the National Academy of Sciences* 109.50: 20425-20430.
- Shariff, SM, M Sanderson, and X Zhang. 2016. "Correlation Analysis of Reader's Demographics and Tweet Credibility Perception." In *Proceedings of the European Conference on Information Retrieval*, pp. 453-465.
- Siebold, GL. 2001. "Core Issues and Theory in Military Sociology." *Journal of Political and Military Sociology* 29(1), 140-159.
- Smith, MC, DA Broniatowski, MJ Paul, and M Dredze. 2016. "Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter." In *Proceedings of AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.
- Soni, S, T Mitra, E Gilbert, and J Eisenstein. 2014. "Modeling Factuality Judgments in Social Media Text." In *Proceedings of ACL*. Baltimore, MD. <http://www.aclweb.org/anthology/P/P14/P14-2068.xhtml>
- Starbird, K, J Maddock, M Orand, P Achterman, and RM Mason. 2014. "Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing." In *Proceedings of iConference*.
- Sueker J, DL Blazes, MC Johns, PJ Blair, PA Sjoberg, JA Tjaden, JM Montgomery, JA Pavlin, DC Schnabel, AA Eick, S Tobias, M Quintana, KG Vest, RL Burke, LE Lindler, JL Mansfield, RL Erickson, KL Russell, and JL Sanchez. 2010. "Influenza and Respiratory Disease Surveillance: The US Military's Global Laboratory-based Network." *Influenza and Other Respiratory Viruses*, 4(3): 155-161.
- Tausczik, YR and JW Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1): 24-54.
- Wang, G, T Konolige, C Wilson, X Wang, H Zheng, and BY Zhao. 2015. "You Are How You Click: Clickstream Analysis for Sybil Detection." In *Proceedings of USENIX Security*, pp. 1-15.

WHO. 2009. *Influenza (Seasonal), Fact Sheet Number 211*. World Health Organization. Available at <http://www.who.int/mediacentre/factsheets/fs211/en/index.html>. Accessed August 2016.

Xia, X, X Yang, C Wu, S Li, and L Bao. 2012. “Information Credibility on Twitter in Emergency Situation.” Chapter in *Intelligence and Security Informatics*, eds. M. Chau, GA Wang, WT Yue, and H Chen, pp. 45–59. Springer: Berlin Heidelberg.

Yang, J, S Counts, MR Morris, and A Hoff. 2013. “Microblog Credibility Perceptions: Comparing the USA and China.” In *Proceedings of the Conference on Computer Supported Cooperative Work*, pp. 575-586. ACM.

Yang, Y and J Eisenstein. 2013. “A Log-Linear Model for Unsupervised Text Normalization.” In *Proceedings of EMNLP*, pp. 61-72. ACL.



Pacific Northwest
NATIONAL LABORATORY

*Proudly Operated by **Battelle** Since 1965*

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)

U.S. DEPARTMENT OF
ENERGY

www.pnnl.gov