The American Soldier in World War II: Text Analytics Potential in Historical Survey Data

Team AI'd Powers: Chandler Crescentini, Colin Lagator, Quinn Moore

Client: Dr. Ed Gitre, Virginia Tech Department of History

11 December 2019

Executive Summary

Throughout World War II, the Army conducted over 200 different surveys covering a variety of topics in order to gain insights into the mind of the "civilian soldier". Each survey consisted of multiple choice and handwritten response sections. Originally, there was a one-to-one match of these two parts, but over time, this connection was lost. The Army researchers did not have an efficient way to process the large amount of text, so their focus was on the multiple choice responses. Our primary goal was to assist our client, Dr. Gitre in reuniting this data using serial numbers that are associated with some surveys. This process was successful for a subset of the data, but others will need to be developed to reunite the whole collection. In addition, we provided Dr. Gitre with a showcase of different natural language processing (NLP) techniques, including Sentiment Analysis, the CUR Decomposition, and N-grams. These analyses provide insights into the free response data for the first time since it was collected three-quarters of a century ago. The analysis of this data yields a unique and novel look into the minds and humanity of soldiers in World War II.

1 Problem Statement

This semester, we worked with Dr. Ed Gitre from the Virginia Tech Department of History. Dr. Gitre is working on a project titled "The American Soldier in World War II". There is data from this project that consists of surveys taken by United States soldiers in WWII. This data has sat in the National Archives for many years, and a large scale analysis of the entire dataset has never been done. The end goal of his project is to develop a website that conveys the historical value of this data to the public. This project is grant funded, and has a large team of historians and researchers working on it from institutions around the country. Dr. Gitre is working with us to accomplish two main goals. First is the goal of reuniting the data. The surveys consist of free response and multiple choice questions. Until recently, the free response answers existed as scans of handwriting. Dr. Gitre has led a crowdsourced campaign to transcribe this handwriting into usable data. Now that these responses are digitized, they can be reunited with the multiple choice answers, which were digitized in the 1970's. When the surveys were taken, there was a one-to-one matching of the answers to these questions. Over time, as the data was moved and transferred, this connection was lost. We aimed to develop a process to reunite these two types of responses. The second goal is to showcase what kind of analysis techniques can be used with this data. In the past, not much work has been done with the free response answers. We focused on developing NLP techniques to work with this free response data. Over the course of the semester, we worked closely with Dr. Gitre and Dr. Mike Hughes to gain insights into the historical aspects of the data. We also worked with Aaron Schroeder and Gizem Korkmaz from the University of Virginia's Social and Decision Analytics (SDAD) during our process of parsing and reuniting the data.

2 Ethical Considerations

One of the most important ethical considerations for this study as a whole pertains to the anonymity of the soldiers at the time of completion. The War Department was interested in the honest thoughts and opinions of the "civilian soldier," one who was drafted as opposed to willingly enlisted. The level of honesty from survey takers is increased through their anonymity. Ong et. al. (2000) [1] confirm this effect that anonymity has. This is important, due to the fact that there were questions pertaining to opinions on commanding officers and superiors. The respondents may have felt a fear of retribution if they said anything negative. As a result, this anonymity was heavily emphasized and adhered to by the designers and administrators of these surveys.

Another ethical consideration that extends to the modern world is the way in which this data, especially the free response answers, is consumed by the public. Many of these surveys deal with sensitive subjects like attitudes towards race, women, and the enemy. Quite frankly, some of the responses on these subjects will be shocking to the public; as societies' attitudes towards these topics have changed dramatically in the decades following WWII. The public must be reminded and warned that these datasets are representative of soldiers during WWII, not necessarily society as a whole, and that it is not accurate to apply current mainstream moralities to the past.

3 Literature Review

In "BERT: Pre-training of Deep Bidirectional Transformers For Language Understanding", Devlin et al. (2018) [2] introduce the Bidirectional Encoding Representations from Transformers (BERT)

model. The paper goes over the model's applications and it benefits as opposed to other language representation models. This paper also details the possible applications of BERT and the nature of the input that is needed to run the model (i.e. tokenization, sentence tagging, etc.).

In "Efficient Algorithms for CUR and Interpolative Matrix Decompositions", Voronin and Martinsson (2017) [3] describe algorithms for the computation of the CUR decomposition. Numerical experiments on this more efficient algorithm demonstrate superior performance compared to other existing techniques. The algorithms described are based on simple modifications to the classical truncated QR decomposition, meaning that highly optimized library functions can be utilized for the implementation. This paper is relevant to the project because it provides the group with simple and efficient algorithms for computing the CUR matrix decomposition which will aid the speed and ease of implementation.

In the "Impact of Anonymity on Responses to Sensitive Questions", Ong and Weiss (2000) [1] discovered that people are far more likely to be honest about negative behaviors when surveys are completed under anonymous conditions. When the survey was completed under confidentiality 25% of respondents admitted to a negative behavior, but under anonymous conditions, 74% admitted to it. This study is relevant to our project because they both deal with anonymity and surveys. A survey is void if the respondent isn't being honest, and the condition of anonymity supplied to the soldiers when completing these surveys leads us to be confident that they were being honest. This in turn, leads to more accurate and useful surveys for statisticians and scientists to use.

4 Project Criteria

In the following section, we establish the set of criteria that characterize our solutions for this project, as well as provide an appropriate quantitative measure for each one. We have three project components: Data Profiling and Reuniting, Data Analysis Prototyping, and Data Visualization Prototyping.

The Data Profiling and Reuniting component consists of our development of a process to reunite the free response and multiple choice answers.

Criteria	**Metric**
Divisibility	We are able to divide the data into at least 2 usable features.
Applicability	The process we use to unite the data can be applied to 90% of the surveys.
Automation	The reuniting process is successful on 75% of the answers to a survey.

The Data Analysis Prototyping component consists of our development of multiple analysis techniques to be used on the data.

Criteria	**Metric**
Speed	Analysis takes under 30 minutes to run on an i7 laptop.
Transformability	100% of the analyses can be transferred to a website.
Applicability	Analysis can be run on 95% of reunited survey data.
Language	Analysis will be conducted in R or Python.

The Data Visualization Prototyping consists of our development of visualizations that can be used in conjunction with our analyses. Additionally, we developed visuals that could be transferred to a website setting, keeping with the goal of our client's research project.

Criteria	**Metric**
Speed	Visualizations take under 5 seconds to load.
Clarity	Visualizations are ranked a 6 or 7 from 1 (unclear) to 7 (very clear).
Interactivity	50% of visualizations should have at least 1 interactive component.
Variety	Utilize at least 3 visualization techniques.
Applicability	Visualizations can be used with at least 95% of cleaned survey data.

5 Selected Solutions

5.1 Data Profiling/Reuniting

1. Create a Reference Document for the Different Data
   We have many different forms of data coming from many places, it will be important to have a firm grasp of what data we have and where it is. We propose to create a reference document to detail the location and structure of all available data.

2. Match Answers Using Leading Multiple Choice Answers
   Some of the free response questions contain a multiple choice question on the same page. This answer was transcribed with the free response answers. It might be possible to unite some of the free response and multiple choice answers. If the free response data contains multiple choice answers, we can use the full multiple choice answers data to narrow down matches with process of elimination.

3. Use Serial Numbers to Find Multiple Choice-Written Response Pairs
   A subset of the surveys in this collection have been equipped with serial numbers by the designers of this study. However, the serial numbers on the two question types are encoded differently. Matching the two question types will work for this subset of the surveys, but we must first find out how the serial numbers stamped on the free responses are related to the ones in the multiple choice files.

5.2 Data Analysis Prototyping

1. BERT

   This is a language embedding model which provides improvements on previous models in accuracy and run-time. The difference between BERT and previous models is the way the text is processed. Previous models processed text from left to right; however, BERT looks from both the left and the right, which provides more context in the representation of language. It is possible to get vector representations of sentences from BERT, and then run other models on those representations. We could use this method on some of the survey free response questions. A deeper description of the above can be found in Devlin et al. (2018) [2]. This article was the first introduction of the model and provides explanations of the structure of the model, its advantages over others, and possible uses.

2. CUR Decomposition

   This decomposition approximates an $m \times n$ matrix $\boldsymbol{A}$ as a product, where $\boldsymbol{C}$ consists of k columns of $\boldsymbol{A}$, $\boldsymbol{R}$ consists of k rows of $\boldsymbol{A}$, and $\boldsymbol{U}$ is calculated via $\mathbf{U} \approx \mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}$:

$$\underset{m \times n}{\mathbf{A}} \quad \approx \quad \underset{m \times k}{\mathbf{C}} \quad \underset{k \times k}{\mathbf{U}} \quad \underset{k \times n}{\mathbf{R}}$$

   The CUR selects rows and columns that exhibit high statistical leverage or high influence from the data matrix. This technique could be used to identify significant words or phrases from an entire survey subset. It was developed as an alternative to Singular Value Decomposition (SVD) and Principle Component Analysis (PCA). While the CUR decomposition is less accurate than SVD, it is easier to interpret (rows and columns of decomposed matrix mean the same as original matrix). Also, using the methods found during the groups literature review [3], the decomposition can be calculated with a lower asymptotic time complexity.

3. N-grams

   These are a NLP technique that looks at the frequency of multiple adjacent words in a document. The "n" in n-gram indicates the number of words that it chains together. This technique could be used to select phrases that were commonly used in the free response answers. Additionally, it could be used to bring more context to summary statistics like term frequency.

5.3 Data Visualization Prototyping

1. R and Python

   The choices for visualizations will not be limited to packages, but just to R and Python. There are many choices for visualizations in both languages. Finding the most appropriate visual for an analysis will happen on a case by case basis.

6 Results

This results section is the deliverable given to our client. It is meant as a showcase of possible analysis techniques that could be used with the entire survey dataset. This section of the report is meant to be viewed by our client and future analysts that work on the project.

6.1 Summary of Project Results

In this paper, we will detail the analysis techniques that we have researched over the course of the semester. All of the analyses focus on natural language processing (NLP) techniques. Some of the techniques have been developed to work with the reunited free response and multiple choice data. We also demonstrate analyses that can be conducted on the free response answers alone. Details on our techniques for reuniting the free response and multiple choice data are also included. This paper is meant to be a reference for Dr. Ed Gitre and analysts that will work with this data in the future. Relevant code that was used to conduct these analyses will be distributed with this document. For these examples, we worked with Survey 106-EH, which asked question about post-war job plans.

6.2 Data Reuniting

One of the primary objectives of this project was to reunite the handwritten free response questions and the individual level multiple choice files. The two components had been separated after collection, since, at the time, the War Department had no way of analyzing such a large amount of handwritten verbatims. One of the study designers had the foresight to attach serial numbers to some but not all sets of the surveys, so that someone might later be able to draw more meaningful conclusions about each survey as a whole. We used these serial numbers to reunite the answers from Survey 106-EH.

In order to reunite the handwritten verbatims and the individual level multiple choice responses, it was necessary to convert one of them using the table shown in Figure 1 below. We decided to convert the "old" serial number stamped on the handwritten files to the "new" one created by the Roper Center, which was included with the multiple choice answer files. Each serial number consists of five digits with the first two representing the "Branch", the third digit represents the separate outfit within each branch, and the fourth and fifth represent the serial number within each outfit. In order to convert between the two forms, only the first two digits need to be changed.

Branch	**Roper Code**	**Orig. Code 1**	**Orig. Code 2**
Infantry	01	1	1,2
Field Artillery	02	1	3
Coast Artillery	03	1	4
Anti-Aircraft	04	1	5
Engineers	05	2	1
Quartermaster	06	2	2
Ordnance	07	2	3
Signal Corps	08	2	4
Transportation Corps	09	2	5
Military Police	10	2	6
Chemical Warfare	11	2	7
Medical Department	12	2	8
Miscellaneous	13	2	9
Air Corps - Flying	14	3	1,2,3
Air Corps - Ground	15	3	5,6,7
No Code or No Data	00	0	0

Figure 1: Serial Number Conversion Table - Survey 106

The table in Figure 1 represents the key for the serial number conversion for Survey 106-EH. It is currently unknown whether or not this scheme applies to other surveys as well. To show how this conversion works, consider the following example using the first serial number from free response Roll 35.

$$\mathbf{28}463 \rightarrow \mathbf{12}463$$

The serial number on the left side is the "old" serial number that is stamped onto the handwritten free response answers and is included in the transcribed .csv files for this roll. The serial number on the right is the "new" one and is included in the parsed answer .csv files, occupying the final

three columns. In order to convert this number, we simply take the first two numbers and match the first number, 2, with Orig. Code 1, and the second, 8, with Orig. Code 2 in the table above. This identifies the branch, which in this case is Medical Department, and is associated with the Roper Code 12. Changing the first two numbers is all that is necessary; as the outfits within each branch and the serial number within each outfit remain the same between the two forms.

A script in R can be used to automate this process and is called `serial_num_converter_106.R`. Within this script, the function `convert_orig_to_roper` takes the serial number as an argument and is used to convert individual serial numbers. The function `convert_serial_file` takes as arguments the data frame returned from reading in the transcribed free response file (e.g. Roll_35_V2.csv) and either the column name as a string or index as a numeric where the serial numbers reside (e.g. "T7" or 10). This function will convert all serial numbers in the file and append a column to the end of the data frame with the converted number.

Some limitations of this method of reuniting surveys can be observed by examining the numbering scheme in Figure 1. For Infantry, there exist two possible values for Orig. Code 2, and for the two categories of Air Corps, there exist three possible values for each. This means that we will potentially end up with two or three "old" serial numbers being converted to a single "new" serial number. For example, both **11xxx** and **12xxx** will be converted to **01xxx**. One potential solution to this issue is to use leading multiple choice question(s) in order to find the potential match. This will require additional manipulation and cleaning of the data beyond its current form, and will likely only be applicable to a small subset of the overall responses. Due to time constraints, this option was not able to be fully explored and implemented.

6.3 Sentiment Analysis

The purpose and use of sentiment analysis is to computationally measure and categorize text data in order to extract attitude and emotions. What our team wanted to accomplish was try to uncover what the overall emotions that the soldiers were experiencing from these anonymous surveys. With an emphasis on humanity and transparency, we wanted to fully translate the viewpoints of these soldiers.

This led us to using many techniques in R's `tidytext`[1] package. This package offers multiple sentiment analysis techniques across different lexicons. A lexicon is a dictionary that maps words to associated emotions. Some of the other lexicons work on a positive/negative sentiment scale, but we wanted more complex analyses. So, we used the National Research Council Canada (NRC) lexicon. With this lexicon, our team was able to analyze the sentiment of the surveys with more complex sentiment, including emotions such as disgust and anticipation. A radar chart is a useful plot that displays the emotions that make up a section of language.

Figure 2 displays a radar chart for a set of survey free responses. The free responses are sectioned by the answer that the individual soldier gave for a previous multiple choice question. This specific radar chart shows the different sentiment levels of post-war plans grouped by combat experience. As seen in Figure 2, the sentiment is similar for each level of combat. It appears that the sentiment for free response questions does not vary much between multiple choice question grouping. The soldiers had similar ratings for most emotions, but there were some nuanced differences.

[1] Documentation for `tidytext` can be found at: https://cran.r-project.org/web/packages/tidytext/tidytext.pdf

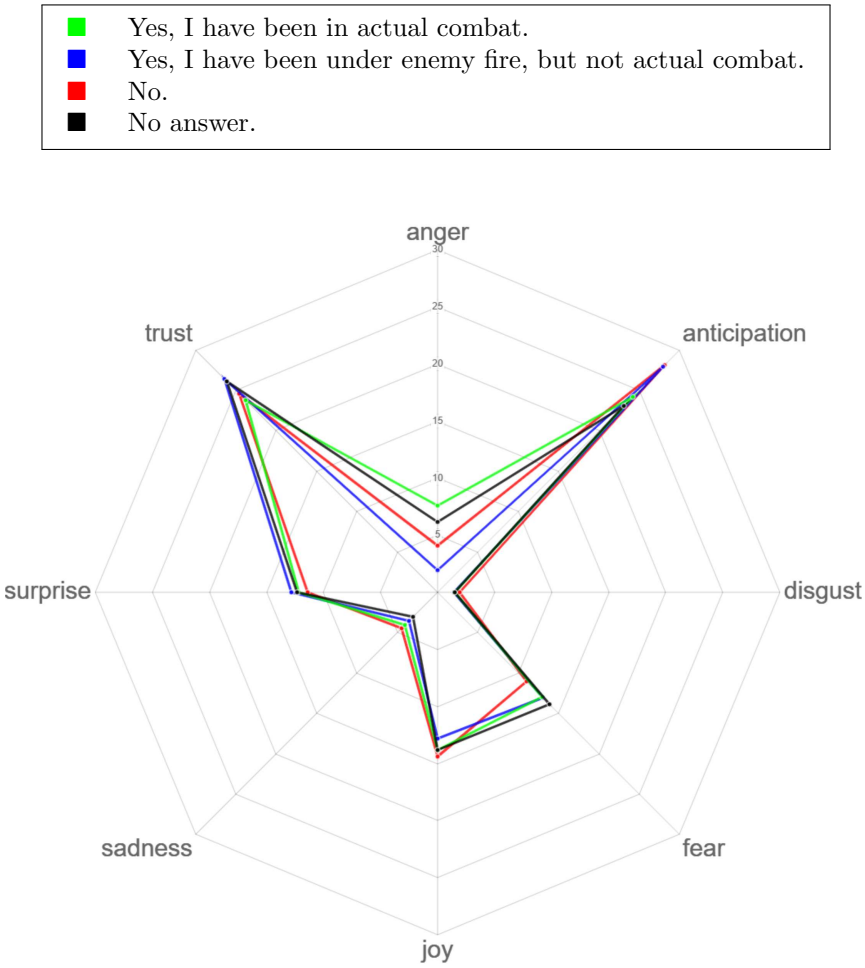

Figure 2: Plot of the sentiment of free response answers grouped by the answers to a leading multiple choice question.

The emotions of interest in this plot would be anger and anticipation. People that have been in actual combat (green), seem to have the most angry responses to the question of post-war plans. Soldiers who have been under enemy fire have greater levels of anticipation in their responses. The ability to catalog the emotions of these soldiers is a powerful tool. Although disgust was not useful in this case, when examining surveys that ask about the enemy, it may be more useful. Sentiment analysis tools like this could potentially help uncover signs of emotional issues with soldiers like post traumatic stress disorder (PTSD). This could be done by using this technique on surveys that focus on mental illness, which was referred to as neural psychosis at the time.

The R code for performing sentiment analysis and radar plots with R's `tidytext` package can be found in the file `sentiment_analysis.R`.

6.4 Bidirectional Encoding Representations from Transformers (BERT)

Bidirectional Encoding Representations from Transformers (BERT) is a model developed by Google for NLP. We are using BERT to get vector embeddings of the free response survey answers. The purpose of creating vector embeddings of language is that we can use numerical analysis techniques that aren't normally applicable to language. There are multiple pre-trained versions of BERT; the version we are currently using outputs a 768 element vector embedding of any inputted language. We created embeddings of the free response answers in Survey 106-EH using this model. These embeddings were used in a CUR decomposition. This decomposition, and its relationship to BERT, will be explained in the next section. All mentions of results in this section are referring to the results of the CUR decomposition.

First, we pre-processed the free responses before inputting them into BERT. The answers are of varying length, some responses are paragraphs, others are just one sentence. At first, we split up all the responses into single sentences, then encoded those. This strategy was not used because we thought longer responses could be over-represented when split up into single sentences. Therefore, we kept all the responses whole. The whole free response answers were cleaned; they were converted to lower case and punctuation was removed. We also removed stop words after noticing that their removal yielded better results. Stopwords refer to common words that don't provide much additional context; in our case, we removed a list of the most common words used in English. We then converted each individual response into an embedded vector; all the vectors were assembled into one matrix. The columns of this matrix represent each embedding. The choice for the columns to represent the data was made because of how the CUR decomposition functions, this will be explained in the next section.

We used `bert-as-service`[2] to run the BERT model. `BERT_106_eh.ipynb` contains the code for this model. `Bert-as-service` is a library available in Python that utilizes pre-trained BERT models. The model that we used was the 12 layer lower-cased English model (uncased_L-12_H-768_A-12 is the official name). When extracting vector embeddings from BERT, it is possible to select the embedding from each layer of the model. For the model we used there are, for example, 12 choices for embeddings. Each layer offers slightly different sensitivity for the context of the language. We looked at the embeddings from the last six layers of the model. We chose the last six because the first six are not very sensitive to context. We felt that this kind of sensitivity was necessary. These six embedding types were looked at to determine the best candidate for the CUR decomposition. We will explain this choice and why the CUR decomposition was used below.

6.5 CUR Decomposition

Our intent with using the CUR decomposition was to utilize low-rank approximation to find a subset of free responses that are most significant. Since there are such a large number of free responses for each survey, we want to be able to present a subset of responses that are representative of the main topics that soldiers wrote about. The classic and optimal way of performing these low rank approximations is with the singular value decomposition. However, this approach results in orthonormal singular vectors that aren't representative of the original data. In order to get an approximation that is representative of the original data, we use an interpolatory factorization known as the CUR decomposition. An explanation of the CUR decomposition can be found in a

[2]Bert-as-service is a Python library that contains a ready to use BERT model, it can be found at https://github.com/hanxiao/bert-as-service.

lecture by Soresnsen and Embree (2016) [4]. This approach returns columns of the original data matrix in the **C** matrix and rows of the original data matrix in the **R** matrix. Since our original data matrix, **A**, is built using the BERT embeddings as column vectors, we are only interested in the **C** matrix.

$$\mathbf{A} \approx \mathbf{CUR}$$

$\mathbf{A} \in \mathbb{R}^{mxn}$, the original data matrix with BERT embeddings as column vectors
$\mathbf{C} \in \mathbb{R}^{mxk}$, subset of the columns of A with k chosen as the number of responses to return
$\mathbf{U} \in \mathbb{R}^{kxk}$, optimizes the approximation
$\mathbf{R} \in \mathbb{R}^{kxn}$, subset of the rows of A with k chosen as the number of responses to return

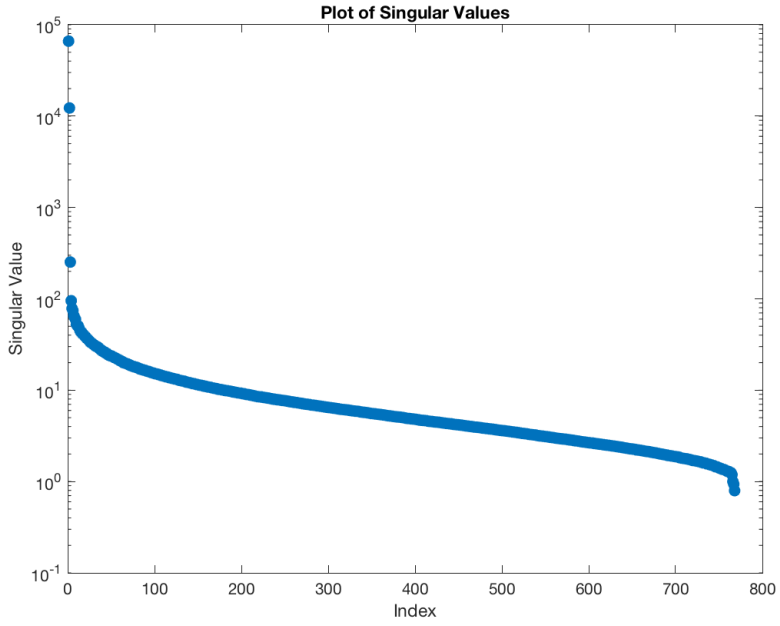

Figure 3: Plot of singular values for BERT encoding 3

Before performing the decomposition, it was necessary to look at the singular values of the original BERT embeddings matrix. We were looking for a specific pattern in the singular values. This pattern is a few large values followed by a significant drop off by a few orders of magnitude. In other applications like image data, we might expect a complete drop off to zero. However, for text data, we will not see this behavior due to the higher levels of variability. Using Python, we computed the singular value decomposition of our matrix of vector embeddings, and then we plotted the singular values. Figure 3 displays the singular value plot, which shows the drop off of singular values that we expected. The behavior just described is indicative of most of the information in the matrix being represented by a small subset of the column vectors. Out of our six possible embedding matrices from BERT, we chose the matrix that showed this pattern the best.

With the behavior of the singular values confirmed, the CUR decomposition was performed on the data matrix of BERT vector embeddings. As mentioned in the previous section, we attempted

the decomposition using several different forms of the original data matrix, with the best results coming from the whole responses with stop words removed. The original transcribed English representations of the **C** matrix, using the free responses from Survey 106-EH, are shown below. During the transcription of the free responses, when the meaning of handwriting is unclear, the words are replaced with tags. Examples that contain unclear tags, which obscure complete context, have been removed. For example "I would like to be a [Unclear]" and "I think every [Unclear]" have been removed from the following list.

- I intend to go back to my old job untill I can get a bussiness of my own ready to start. I will need to get machines and material. Also will have to get a loan to get started.
- My plans for after the war is to work for my factory, and in spare time do carpenter work.
- I am undecided
- All men who are qualified should have first chance with Air Lines.
- I wanna get home right now- this instant.
- My plans will depend upon conditions in the States. By conditions I mean all of
- I plan to have a wheat & poultry farm combined.
- Just aint got much to say it
- Get married and build a home and raise a family.
- If she's still waiting I may marry yet but old age is creeping up.
- WOULD LIKE SOME COLLEGE EDUCATION THEN A JOB EITHER IN THE U.S.A OT SOME FOREIGN COUNTRY
- Why has the 7th AAF not reassigned any of their ex-combat personnel while the Air Force are doing so

On its own, this technique reveals valuable insights into some of the different thoughts and opinions of the soldiers as a whole. For this survey that deals with postwar job plans, we can see that factory work and farming is of interest to some soldiers, while settling down and getting married is of interest to others. To take it a step further, it would be possible to use the responses returned from the decomposition to establish categories in which to separate all other responses. For example, we know that factory work is representative of a large portion of the responses due to its large singular value which leads to a high placement in the **C** matrix. We could use a text search strategy to pull all the responses that reference working in a factory and use the reunited multiple choice answers to perform more analyses on this subset. Alternatively, the reunited multiple choice answers to questions such as combat experience or job in the military could be used to subset these responses and perform the decomposition on these subsets.

The R code for computing the decomposition can be found in the file `cur_106.R` and utilizes the `rsvd`[3] package and the `rcur` function within this package.

[3]Documentation for `rsvd` can be found at: https://cran.r-project.org/web/packages/rsvd/rsvd.pdf

6.6 N-grams

N-grams are similar to term frequency statistics. However, instead of counting single word frequency, n-grams count the frequency of multiple adjacent words. The n in n-gram implies the number of words that are chained together. It could be said that regular term frequency is a 1-gram. We are using n-grams to pull out common phrases from the survey responses. The code for creating the n-grams in this section is found in `n_grams_106_eh.R`. Figure 4 below shows the most common n-grams in the survey responses. The frequency of the phrases have been scaled against the other n-grams of the same length. Before finding the n-grams, we cleaned the free responses by converting them to lowercase, removing punctuation, and removing stopwords.

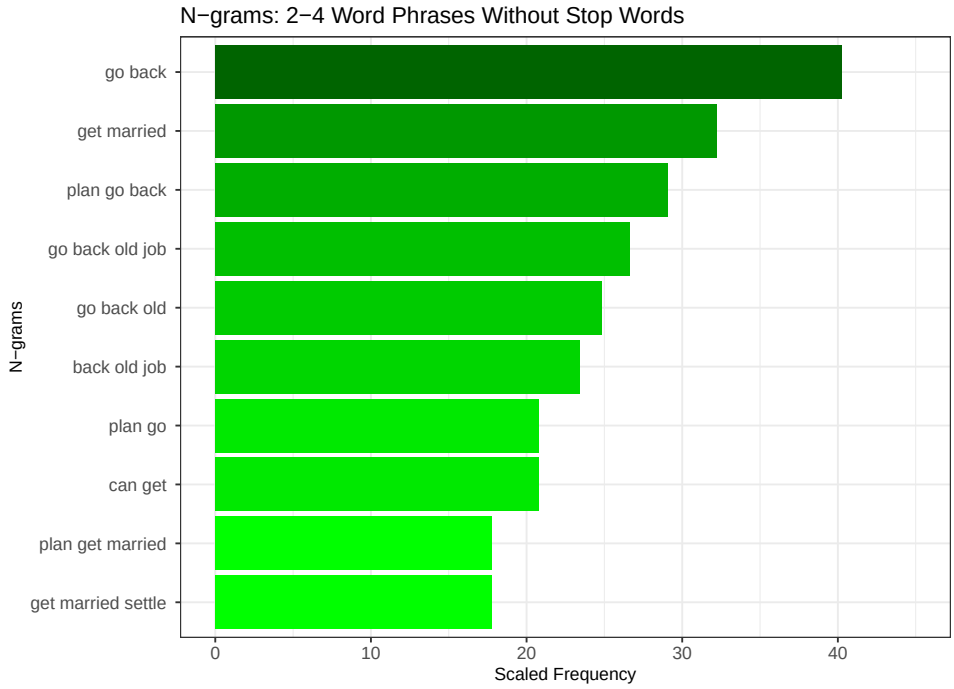

Figure 4: Plot of the most common n-grams

The survey represented by Figure 4 asked soldiers about their post-war job plans. The scaled frequency shows how common each statement was in all the survey free responses. To get the scaled frequencies, the frequencies of the n-grams were normalized. Each group of n-grams, with n ranging from 2 to 4, were normalized with respect to themselves. The scaled frequencies have a mean of 0 and a standard deviation of 1. As a result of this scaling, they can be compared together. We did this scaling because longer n-grams will be naturally less common than shorter n-grams. We implemented the n-gram technique in R using the `ngram`[4] package.

It can be seen in Figure 4 that there are a lot of repeated phrases. This repetition shows a few common subjects, but does not yield much insight. As a result, we further organized the n-grams.

[4] `ngram` is a R package that contains code for creating, displaying, and summarising n-grams. It can be found at https://cran.r-project.org/web/packages/ngram/ngram.pdf.

First, we assembled a list of some of the most commonly used and interesting words from the free responses. This list contained the following: time, life, family, college, home, army, business, married, job, and work. We selected these words from a list of the 25 most used words in the free responses. They were chosen for their specificity to a specific subject. Words like "get" were also common, but not included. The words chosen can be treated as the most popular subjects in the free responses. For each of these words, we found the subset of n-grams that contained it. The n-gram with the highest scaled frequency was selected for each word. In this case, we used the free response data that included stopwords; otherwise, it was cleaned the same. In Figure 5, these n-grams are shown.

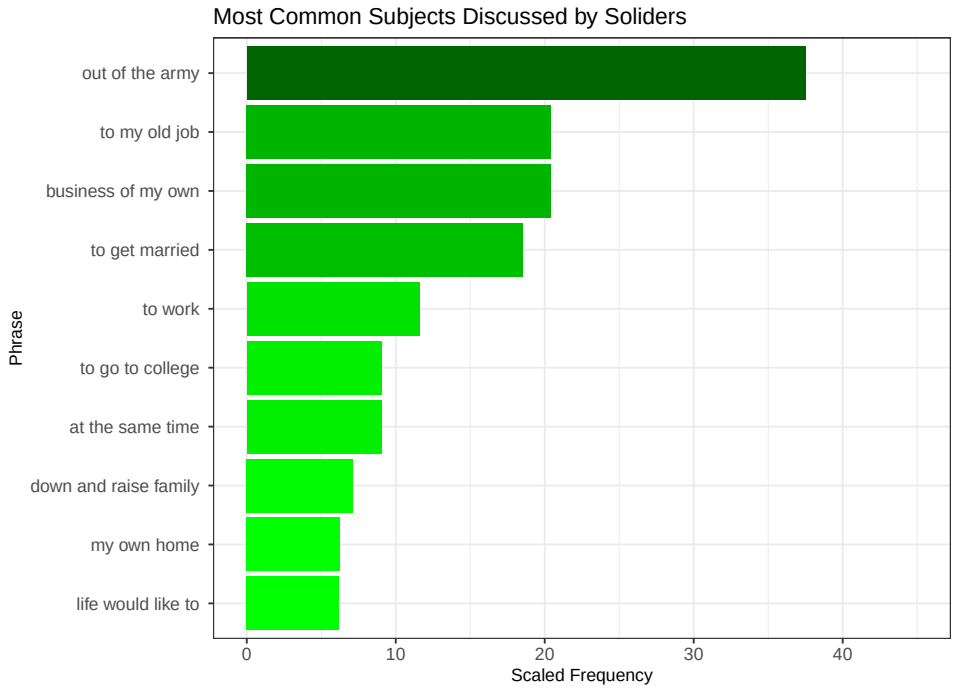

Figure 5: Plot of the most common n-grams used to describe common subjects in the free response survey questions. The frequencies of each type of n-gram have been scaled against themselves.

Figure 5 shows the most popular phrases used to describe the most common subjects in the free responses. It can be seen that a large majority of soldiers talked about getting out of the army. Another insight is that a similar number of soldiers talked about going back to an old job and starting a business. Some of the phrases, like "at the same time", could use more context. However, it can be imagined what solider may have been describing with that phrase, such as working and going to school simultaneously. A phrase like "life would like to" shows that soldiers might have been reflecting on what they want to accomplish in the rest of their life. This plot gives a better idea of the popularity of certain subjects relative to one another, and it is a much more effective use of n-grams than Figure 4.

7 Obstacles

The most significant issue that our team encountered involved the data reuniting phase of the project. The survey data is very disjointed, and it took a great deal of time and meetings at the beginning of the project in order to get accustomed to its structure. Additionally, this is a rather large project with many people and moving parts. It took some time for the group and the client to settle on a particular survey. We needed a survey that satisfied our requirements of having a large number of samples with serial numbers, and which had been transcribed in a usable manner. Furthermore, it took time for the conversion key to be discovered from the large number of historical documents in possession of the research group. After this stage of the project, we were able to move forward with our analyses. However, due to the amount of time spent on the aforementioned stages, we were forced to reevaluate our goals and reduce the scope of the surveys analyzed.

8 Roles

In this section, we will break down the roles that each member of our team had over the course of the semester. Keep in mind that a large portion of work was done collaboratively. However as the project moved forward, especially towards the visualizations phase, each member took on a specialization in order to produce good results.

Colin Lagator:

- *Data Parsing:* The survey data consists of two types, multiple choice and free response. The multiple choice data had been parsed by another member of our client's research group. This data was a rough extraction of information needed from PDF files. At this stage, the data needed to be further parsed and cleaned to align the multiple choice answers with their detailed descriptions.
- *BERT:* Our team planned to use BERT to get vector embeddings of our text data. This required learning how to use BERT and getting a working model running. These BERT embeddings were then used in the CUR decomposition.
- *N-grams:* We make use of n-grams with the text data. In our use, we demonstrate varying lengths of n-grams. In this analysis, we devised a way to compare these varying n-grams and visualize them effectively.

Chandler Crescentini:

- *Serial Numbers and Data Reuniting:* In order to gain meaningful insights from both the free response and multiple choice survey questions, it was necessary to reunite them first. This was accomplished through serial numbers stamped on individual free responses that were converted to the form found in the multiple choice files.
- *CUR Decomposition:* Through the use of the CUR decomposition, we were able to select a subset of the handwritten responses that seem to be representative of the entire data set. We first had to ensure that the singular values behaved in the correct way, and then we performed the decomposition. We ended up with results that seemed to cover a variety of topics and seem to agree with some of our results from the n-grams section. However, we have not been able to define a quantitative way to tell just how effective this technique actually was.

Quinn Moore:

- *Sentiment Analysis:* We utilized sentiment analysis in order to uncover the emotions that are behind the surveys of the general infantry soldiers. These analyses demonstrate that these soldiers are experiencing a multitude of emotions and that these emotions could be a promising way to show the differences from question to question and soldier to soldier. It also brings the humanity of the soldiers to the forefront for people to see.

9 Conclusions and Future Work

Ultimately, the final deliverable of our client's research project is a website for the public to explore these World War II surveys. The results that our team have produced serve as a showcase of the potential for these surveys. In the future, the techniques we've demonstrated may be implemented on the website. Additionally, there is more work to be done to better unite the multiple choice and free response datasets. One option is to use the multiple choice questions found on some of the free response pages. These can be cross referenced with the answers in the multiple choice files to find matches. This will require a rework of the structure of the data, and it will not be applicable to all surveys in the collection. Additionally, it is possible to determine if the data has been collected and stored with some common order. It may be possible to line up free response and multiple choice answers if they have maintained some kind of order.

Future work can build upon our original goals for the project, specifically in the Analysis and Visualizations Scaling section that was ultimately dropped from our workflow. In this section, we planned to scale our reuniting and analysis to other survey datasets. We only had time to work with one survey, but we tried to keep our code open ended, so it could work with other surveys. Other work can focus on applying the techniques we have implemented, as well as others we may have overlooked or have yet to be developed. In the world of natural language processing new techniques are being developed at an extremely fast pace. For example, had this Capstone Project been done in 2018, we would not have been able to utilize BERT in our analysis. The group would like to see the use of the CUR decomposition developed further. Due to the novel nature of this technique in text analytics, there is the potential for this project to to put itself on the forefront of development in an area that is constantly evolving.

References

[1] Anthony D Ong and David J Weiss. "The Impact of Anonymity on Responses to Sensitive Questions". *Journal of Applied Social Psychology*, 30(8):1691–1708, 2000.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers For Language Understanding". *arXiv preprint arXiv:1810.04805*, 2018.

[3] Sergey Voronin and Per-Gunnar Martinsson. "Efficient Algorithms for CUR and Interpolative Matrix Decompositions". *Advances in Computational Mathematics*, 43(3):495–516, 2017.

[4] Dan Sorensen and Mark Embree. "CUR Matrix Factorizations: Algorithms, Analysis, Applications", Nov. 2016, Powerpoint Presentation.