Analyzing U.S. Army Officer Evaluation Reports with Natural Language Processing: A Log-Odds and Latent Dirichlet Allocation Exploration

Heidy Shi, John Caddell, and Julia Lensing
Department of Systems Engineering
United States Military Academy
West Point, NY, 10996, USA

Corresponding author's Email: heidy.z.shi.mil@mail.mil

Author Note: Second Lieutenant Heidy Shi graduated from the United States Military Academy at West Point in 2019 with a Bachelor of Science in Systems Engineering with Honors and a minor in Applied Statistics. Parts of this research were completed as requirements for the Honors Program under the advisement of Major JD Caddell and Major Julia Lensing. 2LT Shi would like to thank the Department of Systems Engineering and Army Human Resources Command for their support of this research.

Abstract: Each job field (branch) in the Army requires a unique set of skills and talents of the officers assigned. Officers who demonstrate the required skills are often more successful in their assigned branch. To better understand how success is described across branches, research was conducted using text mining and text analysis of a data set of Officer Evaluation Reports (OERs). This research looked for common trends and discrepancies across varying branches and like groups of branches by analyzing the narrative portion of OERs. Text analysis methods examined words and bigrams commonly used to describe varying degrees of performance by officers. Topic modeling using Latent Dirichlet Allocation (LDA) was also conducted on top rated narratives to investigate trends and discrepancies in clustering narratives. Findings show that qualitative narratives for the top two performance designations fail to differentiate between officers’ varying levels of performance regardless of branch.

Keywords: Text Mining, Topic Modeling, Latent Dirichlet Allocation, Officer Evaluation Reports (OERs)

1. Introduction

1.1 Background on the Officer Evaluation System and Officer Evaluation Report

Like all enterprises, the U.S. Army is continually seeking the best method for evaluating performance of its officers. The current Officer Evaluation Report (OER) has been a product of decades of research and development and has a significant impact on the careers of all Army Officers. The OER provides necessary feedback to the officer on their performance of duties and provides senior officers the information required to make decisions regarding the officer’s future career (Kite, 1998). These decisions include promotions, assignments, selections for advanced schooling, and retention on active duty (Straffon, 1997).

The newest changes to the officer evaluation system were implemented in 2013. These changes included establishing a distinction between the primary rater and senior rater, altering the performance designations, and implementing a new rater profile to keep track of all OERs completed (Lopez, 2013). Currently, every officer receives two ratings, one from their first-line supervisor (primary rater) and one from the next higher supervisor in the chain of command (senior rater). The primary rater’s responsibility is to evaluate the officer’s performance of duties, based on professionalism, competencies, and attributes. The senior rater’s responsibility is to focus on the officer’s potential for future service and additional responsibility through the ranks. The OER is structured into four blocks where each rater and senior rater can catalogue an officer’s performance compared to others. The rater levels are “excels” which encompasses only the top 49%, “proficient”, “capable”, and “unsatisfactory.” The senior rater levels are “most qualified” which also encompasses only the top 49%, “highly qualified”, “qualified”, and “not qualified.” These block checks (performance designations) are often the first data point examined when evaluating potential to retain or promote officers during a promotion board. These quantitative block checks are paired with a section for qualitative comments where raters and senior raters can input remarks (Department of the Army, 2015).

To avoid over inflation of the evaluation system and to delineate high performers amongst the ranks, the current OER system uses a forced distribution ranking system by limiting the percentage of “most qualified”/“excels” (top block check) that both raters and senior raters may assign. Both raters and senior raters must maintain a “credible” rater profile by assigning less than 50% of the ratings in the top block for any given rank. As part of maintaining a credible profile, raters/senior raters are encouraged to “maintain a cushion” in the number of “most qualified”/“excels” (top block check)
ratings given to prevent exceeding the 50% threshold (Department of the Army, 2015). However, simply selecting a block check may not always tell the full story of an officer’s performance. Ostensibly, the qualitative narrative portion of the report holds merit as well in describing performance and future potential as an officer in the Army. Generally, officer’s feel the senior rater’s narrative has more impact than the rater’s narrative and thus became the focus of this research. Through text analysis this research sought to identify patterns and trends within these narratives.

1.2 Motivation for Research

1.2.1 Limitations to Senior Rater Block Checks

Due to the requirement to maintain a credible profile, senior raters are cautious in the execution of their rankings. Officers with immature rating profiles may choose to reserve top block checks during their initial ratings to build flexibility in their profile. On the other hand, senior raters who mismanage their profile may be unable to reward officers they feel are deserving of a top block rating (Cho, 2015). Inevitably, the system may force the hand of a rater or senior rater to assign a rating that is not a true representation of an officer’s performance or potential and reduces some of the credibility that the block checks should provide.

Additionally, senior raters are unable to annotate if they were unable to give out a top block check because of their senior rater profile limitations (Department of the Army, 2015). As a result, a common assumption is that senior raters often focus on the narrative portion of the OER to try and draw attention to an officer’s potential, despite not being able to award a top block check. Senior raters often put a substantial amount of effort into their senior rater narratives, sometimes choosing to focus on enumeration to try and make certain officers stand out. There may also be certain words that senior raters will use that have more meaning and a stronger effect to portray future potential amongst the officers they rate. This research will test these assumptions to see if there are any trends or discrepancies across OER narratives.

1.2.2 Talent Based Branching

Each branch requires specific talents of its officers and the jobs performed by each officer can vary significantly across branches. Each branch publishes a talent “storyboard” that details specific intelligences, skills, knowledge, and behaviors demanded by each of the 17 basic branches. There are some talents that are in high demand across several branches as well as some heterogeneity across the branches. Table 1 outlines the branch talent demand across the basic branches and indicates that there are closer talent correlations among maneuver branches such as Infantry (IN) and Armor (AR), just as there are correlations among logistics and sustainment branches such as Quartermaster (QM), Ordinance (OD), and Transportation (TC) (Calarusso, Heckel, Lyle, & Skymmyhorn, 2016). Naturally, one would assume there should be discrepancies across the different branches regarding the words senior raters use to describe success and failure amongst Army officers. More interestingly, there may also be common trends across each of the branches that senior raters use to define success. This research will investigate these assumptions to see if there are any trends across different branches.

Table 1. Talent Requirement Matrix by Branch (Calarusso, Heckel, Lyle, & Skymmyhorn, 2016)

<table>
<thead>
<tr>
<th>Talent Demand</th>
<th>AD</th>
<th>AG</th>
<th>AR</th>
<th>AV</th>
<th>CM</th>
<th>CY</th>
<th>EN</th>
<th>FA</th>
<th>FM</th>
<th>IN</th>
<th>MI</th>
<th>MP</th>
<th>MS</th>
<th>OD</th>
<th>EOD</th>
<th>QM</th>
<th>SC</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physically Fit</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mentally Tough</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem Solver</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process Disciplined</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Risk Taker</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatially Intelligent</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technologically Adap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: AD = Air Defense, AG = Adjutant General, AR = Armor, AV = Aviation, CM = Chemical Corps, CY = Cyber, EN = Engineers, FA = Field Artillery, FM = Finance, IN = Infantry, MI = Military Intelligence, MP = Military Police, MS = Medical Service Corps, OD = Ordinance, EOD = Explosive Ordnance Disposal, QM = Quartermaster, SC = Signal Corps, TC = Transportation
1.3 Officer Evaluation Report Data Set

1.3.1 Make-Up of the Data
The data set used to conduct this analysis includes every active duty OER that was submitted with an end date during the 2017 evaluation period. The data set includes over 156,000 entries and is organized based on six factors: gender, branch, rater performance designation (block check), rater narrative, senior rater performance designation (block check), and senior rater narrative. The gender was coerced based from pronouns used within the narrative of each entry. The branches in the data set include all functional areas and 17 basic branches. All data was de-identified to protect confidentiality of the ratees and raters. The rank of the rated officer was also redacted, but the rank for the data set ranged from warrant officers to colonels. Figure 1 is a summary of the total number of OERs per branch broken down by the senior block check rating. This data reflects the relative size and density of officers within each branch. For example, while Aviation (AV) is not the largest branch in the Army, it has a high density of both warrant and commissioned officers as reflected in their 9.2% of total evaluations in 2017.

Figure 1. Total Number of OERS in 2017 by Branch

1.3.2 Limitations to the Data
The narrative portion of the OER is similar to an essay evaluation, which is one of the six categories of methods of performance appraisal often utilized by the U.S. business and industrial sector (Morrisey, 1983). While essay evaluations are excellent for capturing details and providing specific internal feedback, the greatest disadvantage is that they are subjective in nature and are much more difficult to use when comparing to others. Additionally, essay evaluations are limited to the rater’s ability to provide good feedback. Oftentimes, ratees who are evaluated by raters who are better at articulating and writing evaluations may come off as stronger performers than those being rated by raters with a poorer writing ability (Milkovich & Boudreau, 1997). Consequently, focusing on specific words may not be able to accurately tell the full story of how we talk about officer success.

Additionally, the block check portion of the OER is based on a forced distribution rating system to combat inflated ratings and to clearly distinguish high-performing officers from their peers. Proponents of forced distribution rating systems believe that the system motivates individuals, forces raters to be honest with their ratees, and develops strong leaders. It also benefits poor performers by encouraging them to move on to other jobs for which they may be better suited (Welch, 2001). On the other hand, forced distribution systems can also be viewed as unfair, subjective, and vulnerable to biases. It also may discourage collaboration and teamwork (Pfeffer & Sutton, 2000). Thus, focusing on block checks also may not tell the full story of officer success.

Moreover, the current evaluation system limits the data set to provide a perspective of only two superior officers. As a result, these measures only look at mission accomplishment and do not measure molding and motivating soldiers and units for long-term success (Reese, 2002). There may also be other factors that breed officer success that are not described in the
OER. The data set also limits the ability to test the true impact of these evaluations because there is no way to tell whether the officers being rated were actually promoted or selected for promotion or positions of greater responsibility.

2. Methodology

2.1. Exploratory Data Analysis and Data Mining

Exploratory data analysis is defined as the process of discovering interesting patterns and knowledge from large amounts of data. Steps that occur during this process include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. During data cleaning, inconsistent data and noisy data are removed because the data may detract from the data mining process and affect pattern evaluation. Data transformation allows data to be transformed and consolidated into forms appropriate to perform both summary and aggregate operations. Then through data mining, patterns are extracted and evaluated to produce results and conclusions (Han et al., 2011).

Data mining was first conducted on the given data set to ensure the data was usable to produce patterns. The data was cleaned to eliminate all O6 (Colonel) OER narratives, which were filtered based on specific block checks used only in the evaluation of colonels. Empty data fields where no senior narrative was present were also eliminated. Once the data were cleaned, exploratory data analysis was conducted on the new data set to see if the data set was an accurate representation of the true OER profile. Conducting this analysis also helps give a better understanding of the overall data set and allows for discovery of discrepancies to focus on when conducting further research. Figure 2 shows that the overall distribution of OERs given does provide an accurate representation of how senior rater designations (block checks) should be given. There were less than the 50% allotted for “most qualified” block checks, with only 37% of all OERs in the 2017 rating period being given a “most qualified” rating. Most of the OERs were given a “highly qualified” block check, as almost 60% of the OERs were rated “highly qualified.” There were only a small portion of “qualified” and “not qualified” ratings awarded. Oftentimes, these ratings are only given for extremely poor performance or for disciplinary issues.

![Figure 2. Distribution of OERs by Senior Rater Label Block Check During the 2017 Evaluation Period](image)

After understanding the overall distribution of block checks across all OERs given, it is important to compare that distribution across the individual branches to see if block checks across the branches were similar to block checks across the whole Army. Figure 3 shows that across all the basic branches in the given data set that had more than 100 OER evaluations in 2017, the percentage of top blocks given by the senior rater were distributed relatively the same as that of the whole Army. Additionally, none of branches exceeded in assigning more than the allotted 50% of “most qualified” block checks. Each branch stayed well within the 50% threshold, with most OERs given as “highly qualified” OERs. This information suggests that senior raters are maintaining a credible profile regardless of their branch. Moreover, it suggests that officers across the branches are also receiving an equal rating because one branch does give out more “most qualified” block checks than any other branch.
2.2 Text Mining

Text mining discovers and analyzes information, specifically text within documents, to discover patterns. Text mining goes beyond information access to help users analyze and digest information and facilitate decision making. The main goal of text mining is to look for trends and outliers amongst text data. (Aggarwal & Zhai, 2012).

2.2.1 Log Odds Ratio

One method that researchers implement to explore word usage across different documents of text is to use the log odds ratio. The log odds ratio was chosen because it allows researchers to determine words that are more or less likely to come from one specific group of documents. Analysis of OERs was initially done using the common \textit{tf-idf} analysis (Silge & Robinson, 2019). However, \textit{tf-idf} failed to provide the nuance desired. \textit{tf-idf} uses \textit{idf} which is calculated using the equation

\begin{equation}
\text{idf}(\text{term}) = \ln \left( \frac{n_{\text{documents}}}{n_{\text{documents containing term}}} \right)
\end{equation}

In the equation, a term that is used in every document will receive an \textit{idf} of 0 regardless of the frequency of use, and in turn receive a \textit{tf-idf} of 0. Conversely, the log-odds ratio allows researchers to analyze words that may occur in every document or group of documents, and explore to differences based off the relative frequency and uniqueness of use. The log odds ratio can be calculated for each word using equation 2, where \textit{n} is the number of the times the word in question is used and \textit{total} is the total number of words used within each group (Silge & Robinson, 2019).

\begin{equation}
\text{log odds ratio} = \ln \left( \frac{n+1}{\text{total+1}} \right)
\end{equation}

The log odds ratio helps determine which words occur specifically to a single document or group of documents compared to all other words not related to the document or group of documents. Unlike \textit{tf-idf} analysis, words that may occur in every document will still receive a log-odds ratio, especially if one group of documents uses the word much more frequently than the other. Words that are so common that they occur in one group of documents the same amount of times as in the remaining group of documents will have a very low log odds ratio. However, words that are rare for one group of documents will have a
high log-odds ratio. For the OER data set, using the log odds ratio in senior rater comments helps identify words that are frequently used but are more important to groups of documents, such as specific block checks and specific branches. The log odds ratio will pinpoint words that separate the narratives apart for branches or block checks.

2.2.2 Bigrams

Another technique commonly used in text mining is to look at the frequency of consecutive words. Looking for patterns in consecutive words allows researchers to analyze relationships between words. Some common relationships include examining which words tend to follow others immediately or which words tend to occur together across the same documents. Looking at pairs of words also provides more insight to the context of common words used within documents. In order to analyze consecutive words, “n-grams” are used to tokenize adjacent pairs, where “n” is the number of adjacent words being paired together. When looking at consecutive words, researchers will commonly study pairs, also known as bigrams, or triples, also known as “trigrams,” to find relationships and correlations between multiple words (Silge & Robinson, 2019). For the OER data set, analysis of the frequency of bigrams was conducted to gain a better understanding of paired consecutive words used in senior rater comments. The log odds ratio was also applied to bigrams to determine bigrams that distinguished “most qualified” narratives apart from “highly qualified” narratives. Furthermore, a network diagram was built to help simultaneously visualize relationships among words and explore clusters of bigrams.

2.2.3 Topic Modeling Using Latent Dirichlet Allocation (LDA)

Topic modeling is a machine-learning technique that researchers use to uncover patterns within a collection of documents. Oftentimes, documents will contain multiple patterns represented as distributions of words, also known as topics. One type of a probabilistic topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA treats documents as a mixture of topics and topics as a mixture of words. It is a mathematical method used to find mixtures of words associated with each topic while also determining mixtures of topics that describe each document (Silge & Robinson, 2019). LDA allows the researcher to choose how many topics to use to build the model. In turn, the model builds topics and assigns a probability, $\beta$, for each word in each topic. Per-document-per-topic probabilities, represented by $\gamma$, can also be calculated to determine the overall probability that a document comes from a certain topic (Silge & Robinson, 2019). For this research, LDA was used to cluster OER narratives into two topics using bigrams. The data was filtered to only include “most qualified” block checks and “highly qualified” block checks in order to test the hypothesis that the narratives can be divided into its topics by the block check assigned. Ideally, the bigrams used within the “most qualified” narratives should classify the document into one topic and the words used within the “highly qualified” narratives should classify the document into another topic. In doing so, how well narratives were clustered can be observed by comparing the cluster assigned using LDA with the original block check given.

3. Results

3.1 Text Mining for Trends Within Branches

Text mining was conducted on the OER data to look at the log-odds ratio across senior rater comments. Figure 4 expresses the results of text mining senior rater comments across the maneuver branches (Armor, Infantry, and Field Artillery) and the logistics branches (Ordnance, Quartermaster, and Transportation). The top 10 words with the greatest log-odds are shown for each branch. Naturally, each branch has specific words that senior raters use to talk about the officers they rate. For example, common words used by Armor officers include “armor” and “tank”, whereas words used to describe Infantry officers include “rifles,” “mortar,” and “infantry.” When looking at the logistics branches, Quartermaster focuses on other aspects of supply; hence “pbo” (property book officer), “food” and “petroleum” are some of the top words. Transportation also yielded special words across its narratives to include “watercraft” and “vessel” which accurately describes the roles Transportation officers have. The log-odds ratio for each word sorted by the branches helps identify the words that are important to each of the branches amongst the collection of OERs. It shows that these words are what distinguishes the OER narratives apart across each of the different branches.

The words identified show that the words senior raters are using are reflective of the jobs the rated officers have held. These words do not describe attributes that they have demonstrated for their branch. Specific adjectives identifying talent and potential also do not appear. Additionally, since log-odds ratio identifies unique words specific to each branch, the pattern suggests that specific adjectives and descriptors that are used may be so common across all branches that it is not reflected at all when analyzing OERs. As a result, the only words used by senior raters to differentiate narratives across the branches are words indicative of past jobs held or possible future job opportunities.
3.2 Text Mining by Bigrams

3.2.1 Text Frequency

Text mining using bigrams was also conducted to look at patterns of consecutive words used in OERs. Figure 5 displays four graphs of the most frequent bigrams across the maneuver branches and the logistics branches, filtered by two block checks, “most qualified” and “highly qualified.” By analyzing these bigrams across the “most qualified” and “highly qualified” narratives, patterns can be discovered to see if the same words by senior raters are being used, regardless of the block check given.

Across the maneuver branches, one of the most common bigrams used to talk about “most qualified” officers is “unlimited potential.” This bigram is a key bigram that most senior raters use as the frequency of the bigram is significantly higher compared to that of other bigrams. Additionally, senior raters are frequently writing “promote ahead” which is consistent with the meaning of giving out “most qualified” block checks. Looking at the bigrams used in “highly qualified” ratings, it is interesting to note that the common words used include “unlimited potential” and “promote ahead.” The bigrams used in “most qualified” and “highly qualified” narratives are very similar, which may indicate that the narratives between the block checks given are very similar, regardless of the block check given. Ultimately, this could result in two different possibilities. One, officers within the maneuver branches may not be distinguished from one another by the narratives given, and only by the block checks provided. Second, officers within the maneuver branches are using the same words in the narratives given because they are unable to give out the top block check and are attempting to compensate with the narrative. Either way, both cases are indicative that there is a decrease in the relative importance of the narratives, serves as an attempt to circumvent the delineation process, and overall makes the OER a less effective evaluation tool for the Army.

Across the logistics branches, the most common bigrams used to talk about the “most qualified” officers are also “unlimited potential” and “promote ahead.” “Promote ahead” and “promote immediately” are also common words, which indicates that the words senior raters are using to talk about their “most qualified” officers are consistent with the intent of awarding the top block check. When looking at the bigrams used in “highly qualified” ratings, it is interesting to note that once again, the same words are being used in “most qualified” and “highly qualified” OER narratives. Similar to the maneuver branches, the logistics branches reveal the same pattern between “most qualified” and “highly qualified” senior rater narratives. Furthermore, the words used in both the logistics and maneuver branches are almost identical. This pattern may indicate that there is a common standard that senior raters use when constructing their top block narratives.

---

Figure 4. Greatest Log-Odds Ratio on Words in the Senior Rater Narrative for Maneuver and Logistics Branches
3.2.2 Network of Bigrams

A network diagram constructed of different bigrams provides additional insight to how terms are organized and the relationships amongst the words. The network diagram in Figure 6 was constructed using bigrams for both “most qualified” and “highly qualified” narratives on maneuver branches. The diagram helps identify clusters of bigrams that commonly occur together. The width of each edge in the network diagram is determined by the term frequency, where higher term frequencies have a wider edge. The size of each vertex is also weighted based on the degree, the number of edges connected to each vertex. Consistent with Figure 5, “unlimited potential” occurs much more frequently compared to other bigrams within the network as evident by the width of the edge. One interesting cluster is the bigrams generated using “top,” as it has the largest vertex for both groups of narratives. Within the “most qualified” narratives, “top” is followed by various numbers, ranging from “1” to “20” with “5” and “10” occurring more frequently. Within the “highly qualified” narratives, “top” is followed by various numbers as well, ranging from “5” to “50”, to include “half”, with “10” occurring more frequently. It is interesting to note the overlap that occurs within numbers ranging from “5” to “20.” Officers that receive “top 5” within their narratives may fall in either a “most qualified” or “highly qualified” block check, which may be indicative of the narrative compensating for the block check. However, there is still some disparity as expected in that narratives that include “top 1” or “top 2” are rated as “most qualified” narrative whereas narratives that include “top 50” and “top half” are rated as “highly qualified” narratives which are indicative that the narrative aligns with the awarded block check. When analyzing the logistics branches, similar results were produced.
3.2.3 Log-Odds Ratio on Bigrams

While analyzing the most frequently used bigrams provides some insight into how top-rated narratives are structured, analyzing the log-odds ratio using bigrams can help extract bigrams that are specific to each type of branch and type of narrative. Figure 7 shows the top ten greatest log-odds for bigrams in “most qualified” and “highly qualified” narratives for both maneuver and logistics branches. Within the maneuver branches, bigrams that stand out for “most qualified” narratives include “tactical brigade” and “promote bz” (below zone) which are very specific words indicative of a high-performing officer, deserving of a most qualified block check. Bigrams that stand out for “highly qualified” narratives include “soldier unavailable,” “solid performance,” and “solid potential.” These bigrams are indicative of an officer deserving a “highly qualified” block check as they are bigrams indicating “solid” performance as compared to excelling ahead of peers. Across the logistics branches, similar patterns occur. Top bigrams for “most qualified” narratives include “strategic thinker” and “promote bz” whereas top bigrams for “highly qualified” narratives include “solid performance” and “strong candidate.” Using the log odds ratio helps pinpoint bigrams that are distinct to each type of narrative. Thus, it is possible to recognize that officers receiving “solid” as an adjective to describe their performance will have a high chance of receiving a “highly qualified” block check and officers that receive “promote bz” will have a high chance of receiving a “most qualified” block check.

Across the different types of branches, it is also promising to see that narratives for “most qualified” maneuver branches focus on being “tactical” whereas “most qualified” logistics branches focus on describing officers as “strategic thinkers.” This shows that there is some indication of different types of skill sets that are being addressed within the “most qualified” narratives for different branches. However, within the “highly qualified” narratives, the bigrams are very similar, using “solid” as a very common adjective to describe performance, regardless of branch.
3.3 Topic Modeling Using Latent Dirichlet Allocation

LDA was performed to cluster the top-rated narratives by its bigrams. LDA was conducted by breaking down narratives into two different topics to see how well the LDA model was able to classify the narratives into the two different topics. The goal was to see if the model could build two topics that correlated to the block check for that narrative. The model calculates the probability ($\gamma$) that a document is composed of each of the two topics. For example, a narrative could have a probability of being 0.8 of topic 1 and 0.2 of topic 2. In this case, the dominating topic for that narrative would be topic 1, which then classifies the narrative as belonging to topic 1. This classification was completed on all maneuver branch narratives to determine under which topic the model prescribed to the narrative. Figure 8 shows the breakdown of the prediction using the LDA model compared to the original distribution of top-qualified narratives. The actual distribution (or baseline model) had 43% of the narratives as “most qualified” narratives with the remaining 57% classified as “highly qualified” narratives. The figure shows that the LDA model predicted a 7.2% increase in “most qualified” narratives (coerced from topic 2) compared to the original distribution. This shows that a portion of “highly qualified” block checks have been classified as “most qualified” block checks instead, indicating that some “highly qualified” narratives may be written similar to “most qualified” narratives, which caused the model to predict the narrative as a “most qualified” narrative.
Closer analysis through a confusion matrix showed a comparison of how well the model performed in classifying narratives. The matrix in Table 2 reveals that the model was able to correctly predict the block check assigned for 10,577 narratives and incorrectly predicted the block check assigned for 10,033 narratives, giving the model a 51.3% accuracy. The accuracy of the model indicates that the model did a poor job of predicting the correct block check for each narrative. Almost half of the time, the model put “most qualified” narratives as “highly qualified” narratives and “highly qualified” narratives as “most qualified” narratives. This shows that the model could not distinguish the two different topics well. Subsequently, the narratives across the top-rated branches may be written very similar to where a machine-learning algorithm has difficulty distinguishing two groups, or topics, based on the block checks.

### Table 2. Confusion Matrix on LDA Model

<table>
<thead>
<tr>
<th>Reference</th>
<th>Highly Qualified</th>
<th>Most Qualified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Qualified</td>
<td>5993</td>
<td>4275</td>
</tr>
<tr>
<td>Most Qualified</td>
<td>5758</td>
<td>4584</td>
</tr>
</tbody>
</table>

Analyzing the per-document-per-topic probability shows how each narrative was clustered. Each document probability was derived from its bigram-topic probabilities (β) predicted from the LDA model. Figure 9 illustrates the bigrams that had the highest difference in β between the two topics across the maneuver branches. The bigrams with the greatest difference in β are bigrams that are specific to each of the topics. The figure shows that words more common in topic 2 include “ranks 1” and “potential select” whereas bigrams more common in topic 1 include “battalion commanders” and “operations officer.” From previous analysis and the figure below, it is evident that the model was able to distinguish two different topics. However, it seems that the topics are not correlated to the block checks given. The bigrams indicate that topic 1 may be more related to possible jobs whereas topic 2 may be more related to promotion and potential. The breakdown of these bigrams by topics gives an indication of how the model classified each of the narratives into its respective topics. Additionally, it provides further insight in showing that while the model was able to cluster narratives into two different topics, it was not based on the block checks but a different topic coerced from how narratives are written. As a result, it is possible that it is difficult to distinguish how senior-rated narratives are written based on its block checks.
4. Future Research and Conclusion

The 2017 OER data set provides many opportunities to conduct further research to gain a better understanding of how raters speak about officer performance and potential. Further research can analyze primary rater narratives to analyze trends and inconsistencies when evaluating officer performance. Research can also be conducted to compare the narratives of primary raters to senior raters to see if there are patterns or discrepancies between what primary raters and senior raters are saying about their top-rated officers. Additionally, it may be worth confirming raters’ consistency with Army guidance on how to write OERs as well.

This research sought to better understand the importance and impact of narratives on the Army OER through the use of natural language processing. Across the branches, senior raters are using very specific words relevant to their own branches to talk about the successes and failures of other officers. However, these specific words only focus on jobs previously held or recommendations for future jobs that officers may hold. Descriptive words focusing on talent and skills that should be different for each branch are not present and do not align with the Army’s intent of stressing talent-based branching. More interestingly, when analyzing text frequency, officers use similar language to describe their “most qualified” and “highly qualified” officers across maneuver and logistics branches. Using LDA modeling to cluster narratives into two distinct topics failed to align narratives to rater block checks. The model yielded low accuracy due to the similar styles “most qualified” and “highly qualified” narratives were written. This implies that officers across the branches have adapted a common standard to talk about officers in their narratives, irrespective of the block check or the branch. Additionally, this research suggests that the narrative portion fails to make significant distinctions between its officers. As a result, the narrative portion of the OER carries less importance when compared to block checks and may be irrelevant in evaluating officer performance and potential.

5. References

There are no sources in the current document.