Describing Success by Gender Through the U.S. Army Officer Evaluation System

Ellie Senft, John Caddell, and Julia Lensing

Department of Systems Engineering
United States Military Academy
West Point, NY 10996, USA

Corresponding author's Email: ellie.senft@westpoint.edu

Author Note: CDT Ellie Senft is a senior at the United States Military Academy at West Point majoring in Engineering Management. This research was completed as part of the requirements for the Honors Program under the advisement of John Caddell and Julia Lensing. John Caddell is a Major in the U.S. Army and is currently serving as an Instructor in the Department of Systems Engineering. Julia Lensing is a Major in the U.S. Army and is currently serving as an Assistant Professor in the Department of Systems Engineering. CDT Senft would like to thank the Department of Systems Engineering and the Army Human Resources Command for their support of this research.

Abstract: The United States Army uses both subjective and objective evaluation methods when assessing the performance of duties and potential for future service in the Officer Evaluation Report (OER). Males and females proportionally receive the same objective ratings, but on the surface, it is difficult to determine whether subjective ratings are equal. This paper seeks to examine the different ways success is described in each gender and how the OER follows or deviates from these trends. Upon examination of narratives written on the evaluation reports, many of the same words are used to describe success of males and females in the narratives written by their raters. The similarities amongst the reports suggest that the narratives follow a standardized format which may devalue their purpose of providing individualized feedback to the officer and to promotion boards.

Keywords: Text Mining, Sentiment Analysis, Gender, Officer Evaluation Reports (OERs)

1. Introduction

1.1 Background Information

Although women have been serving in the military since 1943, it was not until 2015 that they began integrating into combat arms units. Traditionally, women have served in roles such as nurses, seamstresses, and cooks, but over time they have taken on larger roles and joined new sectors. With the ongoing war on terror and complex battlefield operations, it is evident that everyone must be prepared to fight, regardless of their gender. Many women are highly qualified to perform combat arms operations, yet the current culture has been slow to adjust to the changes (Trobaugh, 2018). The integration of women into combat arms units raises the question of whether or women are evaluated the same way men are in their new, unconventional roles.

Gender stereotypes are social beliefs about the characteristics and attributes associated with each sex (Gupta, 2009). In today's society, men are typically described using agentic qualities, such as analytical, independent, aggressive, and courageous. Women, on the other hand, are typically described using communal qualities such as compassionate, expressive, kind, and supportive (Smith, 2018, Gupta, 2009, Eagly, 2018). Working women are often described as honest, ethical, innovative, and ambitious (Parker, 2015).

Studies show that when men and women are rated objectively, such as through grades, fitness test scores, or class standings, there are no substantial differences in the genders. However, when it comes to subjective ratings, findings suggest that women tend to receive a larger, more diverse set of negative attributes than men, to include being described as inept, selfish, passive, and vain. (Smith, 2018). Many claim that authoritative characteristics typically in males are perceived as more desirable than the communal characteristics in females (Smith, 2019). When women step outside their conventional behavior and act with favorable qualities typically seen in men, they can be criticized for acting outside their gender role. This study seeks to analyze if women in the Army face similar challenges on their evaluation reports.

1.2 Evaluations in the U.S. Army

In the U.S. Army, officers are evaluated at a minimum of every year using the Officer Evaluation Report (OER). The form outlines the duties and responsibilities of each officer, evaluates their performance by their primary rater, and assesses their potential by their senior rater (DA Form 67-10-1). The OER is used for centralized selection and promotion boards, assignment and retention considerations, and professional development opportunities (Ecklund 2006, Kite, 1998). The evaluation is meant to measure performance as well as potential for future service and focuses on leadership capabilities (Ecklund 2006, Hardaway, 2008). The report communicates a recommendation to an officer based on attributes and competencies such as their character, presence, intellect and ability to achieve outcomes (Kite, 1998, United States Army Human Resources Command, 2014). By having officers' complete self-assessments prior to receiving an OER, they are able to reflect on their own leadership and accomplishments (Fallesen, 2017).

There are six sections of the OER analyzed in this study, to include an officer's gender, branch, primary rater label, senior rater label, primary rater narrative, and senior rater narrative. An officers' primary rater is their first-line supervisor, who rates their performance as either "Excels," "Proficient," "Capable," or "Unsatisfactory," followed by a corresponding narrative. A senior rater is an officers' supervisor two levels higher, who rates their potential as "Most Qualified," "Highly Qualified," "Qualified," or "Not Qualified," and also provides a subjective narrative. The following analysis focused exclusively on senior rater labels and narratives because these are what typically hold the most weight when assessing an officer in a promotion board.

OER rater labels are force distributed to ensure differentiation in the ratings, ostensibly to help the Army identify the superior performers within an officer's peer group. A senior rater's profile, or the proportion of officers they rate "Most Qualified," cannot exceed 49%. Senior raters are limited in the number of "Most Qualified" ratings they can give out to officers and will often remain below this threshold and reserve slots to award to the deserving officers in their unit. Additionally, small rating pools naturally force this percentage to fall below 49%. For instance, if a rater ranks six officers, to remain under 49%, they can select at a maximum only two out of the six to be rated "Most Qualified," amounting to just over 33%. There is no cap on the number of "Highly Qualified" ratings that can be given out by raters, which explains why a majority of officers receive this label. Figure 1 shows the breakdown of senior ratings for male and female officers in the year 2017.

Figure 1. Percent of Total Evaluations by Gender

Figure 1 depicts that the percentage of males and females that are rated as "Most Qualified," "Highly Qualified," "Qualified," and "Not Qualified" are roughly equal. Additionally, the percentage receiving the "Most Qualified" label is

significantly below the threshold of 49%, with the majority of officers rated as "Highly Qualified," showing that senior raters abide by the profile limitations.

2. Data

The dataset being used for analysis contains the 156,178 OERs written on active duty warrant officers through lieutenant colonels in the year 2017. An explanation for the different columns and the values possible for each can be found in Table 1 below.

Table 1. Key Variables in Dataset

Variable Name	Possible Values	Notes
Gender	M, F	The gender of the servicemember evaluated
Rater Label	Excels, Proficient, Capable, Unsatisfactory	The rating given to the servicemember by their primary rater
Rater Narrative	Text	Used for the primary rater to provide commentary on the officer's current performance
Senior Rater Label	Most Qualified, Highly Qualified, Qualified, Not Qualified	The rating given to the servicemember by their senior rater
Senior Rater Narrative	Text	Used for the senior rater to provide commentary on the officer's potential

Although the original narratives include the officer's name, the dataset has removed proper nouns and replaced them with x's to de-identify the data. Additionally, because an officer's gender is not explicitly stated on their OER, this variable was coerced through the use of pronouns within the narratives of each OER. The dataset was filtered to remove 5,962 entries without an identifiable gender.

3. Methods

Sentiment analysis aims to assess the opinion, subjectivity, or polarity of text (Pang, 2008). One way to perform sentiment analysis is to calculate the term frequency, or how frequently a word occurs in a document. The individual words in the OER narratives were counted to generate a relative frequency for each. In this analysis, two sub-sets of data filtered by gender were created to discover the words that appear most frequently in the narratives written for each gender. Additionally, this same analysis can be performed on bigrams, or two-word combinations, to pull out the most commonly used word pairs as compared to singular words. Bigrams are useful to study the structure of a dataset by examining the context in which certain words are used together (Silge, 2019). For instance, the frequency of the word "performance" does not hold meaning unless the preceding word of "high" or "low" is known in order to understand the context in which it is used.

Another approach to sentiment analysis is called term frequency-inverse document frequency (tf-idf), which focuses on finding words that are used frequently but are not the most commonly used (Silge, 2019). This pulls out words that are important but are not used as regularly in a collection of documents to find rare or unique words that hold significance. The inverse document frequency for a given word can be found using the equation

$$tf - idf(term) = \ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right) \quad (1)$$

Say, for example, there are 100,000,000 documents in a collection, with the word "that" appearing in all of them and the word "hawk" appearing in just 1,000. Using Equation 1, the word "that" would be assigned a value of 0, while the word "hawk" would be assigned a value of 5 (Enge, 2015). The weight for more commonly used words is decreased, while the weight for rarer words is increased. By performing this analysis on an entire dataset, each word can get ranked based on its tf-idf value to identify the unique words that appear in the documents. The tf-idf statistic measures how important a word is to a document in a collection of documents, or in this case a single evaluation in a collection of OERs. This can be helpful to look below the surface and differentiate the unique words from the ones commonly used.

4. Results

After extracting term frequencies for each gender, it was easily determined that the aggregate language utilized matched almost identically. Figure 2 is a frequency plot displaying words most commonly used when describing males and females rated "Most Qualified" in their senior rater narratives. Many identical words, such as "potential," "promote," "top," and "select" are among the most frequently used in the evaluation reports of high-performing male and female officers. This suggests that there are many similarities in the narratives written for each gender.

The words that show up the most frequently to describe women rated as "Most Qualified" are "potential," "promote," "senior," "top," and "select." The word "potential" has the highest frequency among female evaluations with it occurring in 81.67%, or 7,709 out of 9,428 "Most Qualified" OERs. "Promote" was the second most common word, with a frequency of 66.82%, followed by "top" at 60.75% and "select" at 53.03%. The words that show up most commonly in men rated as "Most Qualified" are "potential," "senior," "promote," "top," and "command." The word "potential" appeared in 34,853 out of 43,412 OERs for a percentage of 80.28%, followed by "promote" at 62.97%, "top" at 60.68%, and "command" at 50.67%.

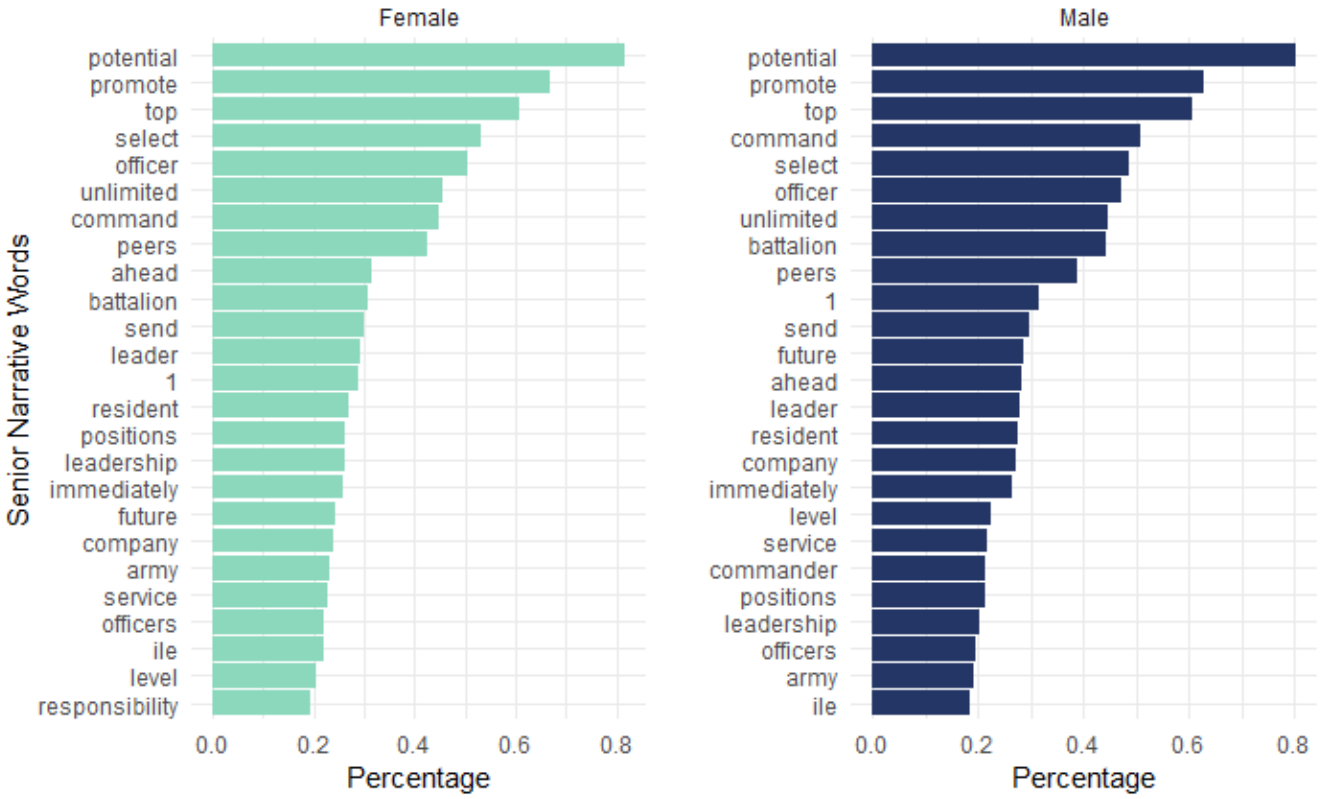

Figure 2. Frequency Plot of Most Qualified Males and Females

The frequency plot shows many patterns between the evaluations of highly successful men and women. The most common word to describe both genders is "potential", but it is unclear without the context whether this refers to high or low potential. The word "promote" also shows up frequently for both men and women because senior raters make judgements about whether the officer deserves to be promoted or not. Due to the finding that the word "top" is among the most common for both genders, it can be deduced that both males and females were rated at the top of the other officers they were evaluated against.

Word-pair analysis using bigrams also revealed many similarities between male and female officers rated "Most Qualified" as seen in Figure 3.

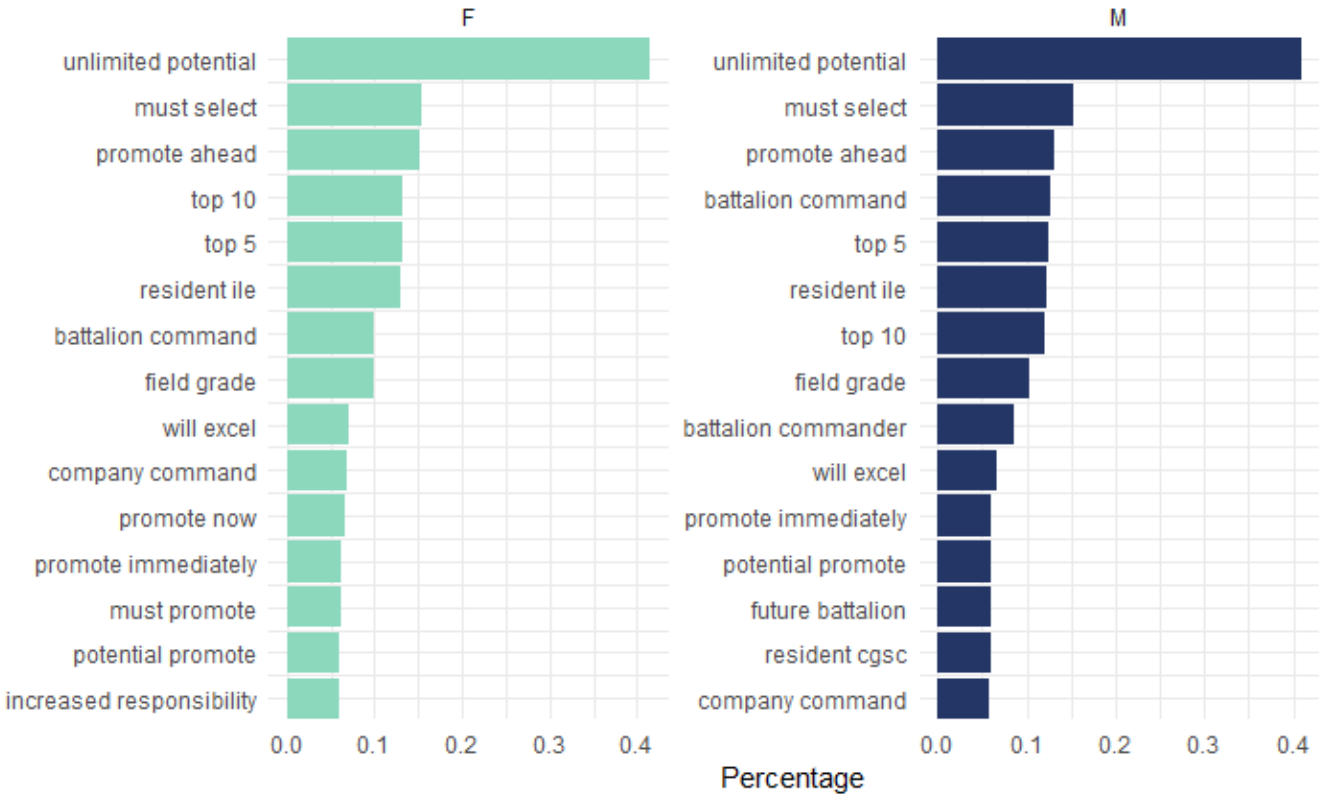

Figure 3. Bigram Frequency Plot of Most Qualified Males and Females

The three most common word-pairs that appear in both OERs include "unlimited potential," "must select," and "promote ahead." Out of the over 43,000 males rated "Most Qualified," 17,755, or 40.89%, included the phrase "unlimited potential," with 41.52% for females. These striking similarities suggests that males and females are proportionally described the same. Additionally, roughly the same proportion of males and females were described as being in the "top five" officers in their peer group. 12.38% of males were described as "top 5," while 13.17% of females earned the same ranking. Similarly, less than 1% more females than males were described as being "top 10." The bigrams display extremely similar word-pairs used to describe males and females used in nearly identical proportions of evaluation reports.

The bigrams are useful because they give context to many words found in Figure 1. For instance, the word "potential" has a high frequency because many officers are described as having "unlimited potential" or being a "potential promote." Additionally, the word "promote" has a high frequency for both genders because senior raters make recommendations on whether an officer should be promoted ahead of their peers, immediately, or potentially. The bigrams help to draw stronger conclusions than the initial frequency analysis by providing context to depict the usage of high-frequency words.

To look beneath the surface and examine more rare words that are commonly used in the evaluation narratives, a plot of the inverse document frequencies can be found in Figure 4. The initial inspection of "Most Qualified" OERs identified a significant presence of nicknames used in the narratives. This is in stark contrast to "Qualified" and "Not Qualified" OERs which maintained more formal language. This suggests that senior raters project a closer relationship to officers they rate "Most Qualified" and use formalities when writing negative evaluation reports. These nicknames were removed from the dataset to allow for additional insights to be made.

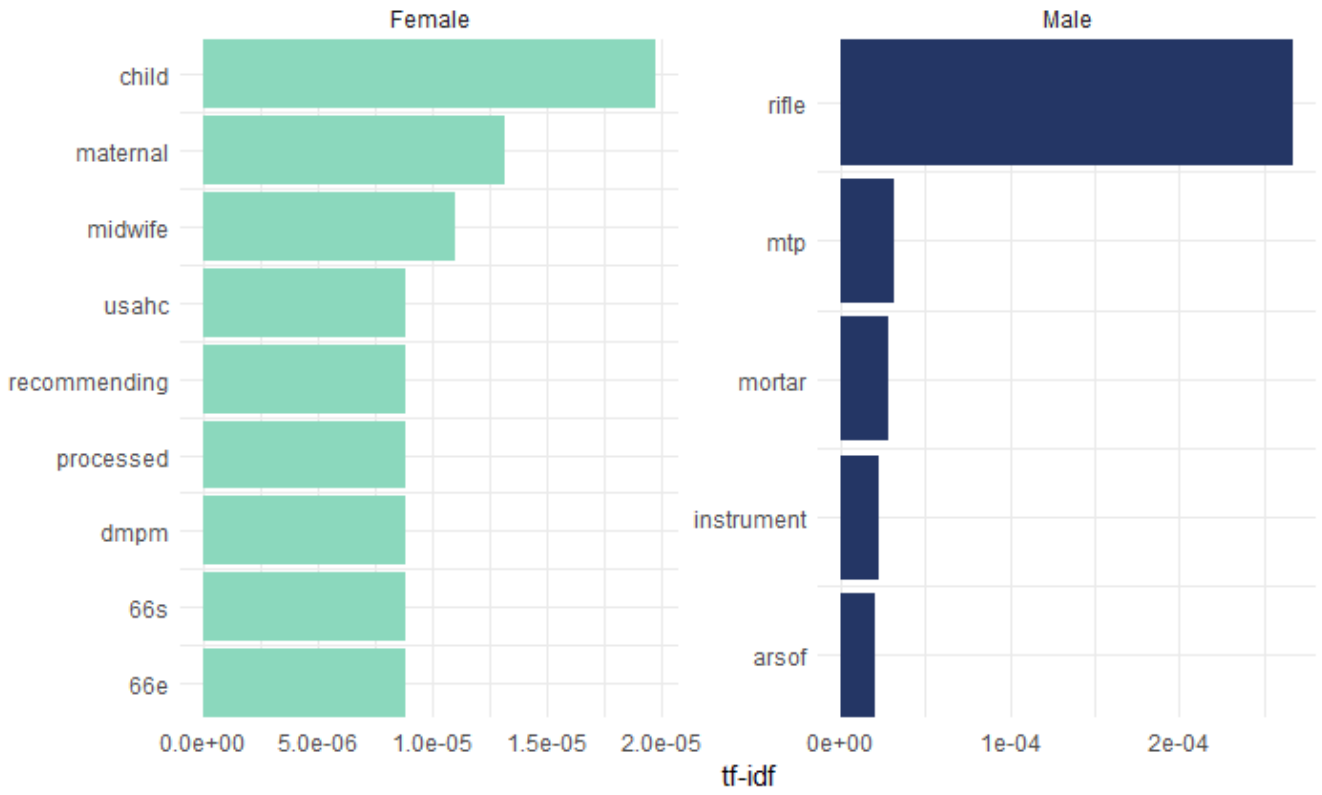

Figure 4. Inverse Document Frequencies for Most Qualified Males and Females

Figure 4 shows that the most common unique words that appear in Most Qualified officers appear to be the specific jobs that the officers hold. More women than men work in the U.S. Army Hospital Corps (USAHC) as a 66E, or a perioperative nurse, which explains why words found for females include "child," "maternal," and "midwife." On the other hand, men are typically in combat arms branches where they are rifle or mortar platoon leaders, and many are also members of the Army Special Operations Command (ARSOF). The inverse document frequency plot shows that the unique words to describe officers are job titles, not unique descriptions about the individual's performance. The extremely small tf-idf values display that there are very few rare words in OER narratives, which suggests that OERs are mostly filled with common words and phrases.

The tf-idf plot in Figure 5 uses bigrams rather than singular words to display the rarer two-word combinations that are frequently used in the Senior Rater narratives.

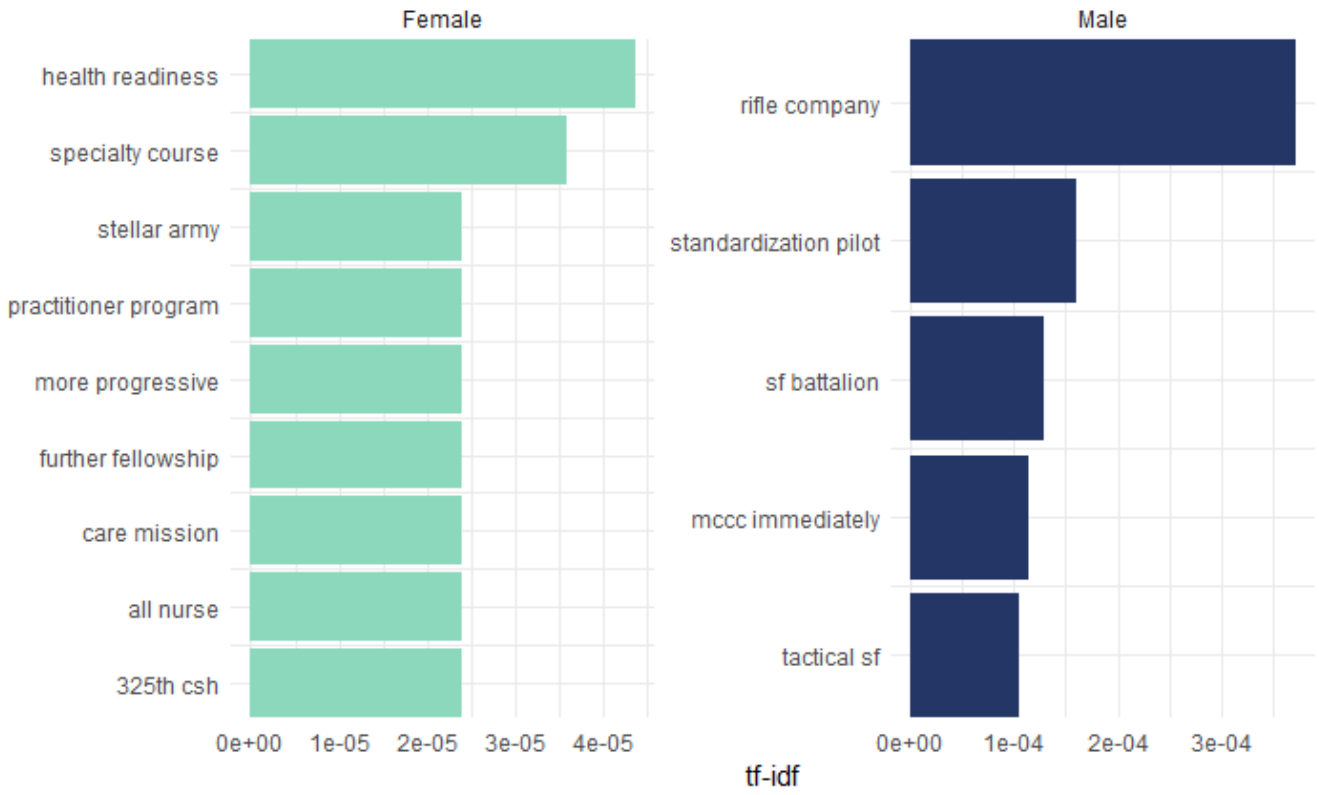

Figure 5. Inverse Document Frequency Bigrams for Most Qualified Males and Females

The findings of the bigram analysis are very similar to those of the tf-idf of individual words, where the words are references to jobs or positions the officer holds in the Army. The bigrams that appear for females include the phrases "health readiness," "practitioner program," "care mission," "all nurse," and the "325th CHS," which is the abbreviation for a Combat Support Hospital. For males, the bigrams include the context for words such as "rifle company," "standardization pilot," "SF battalion," or special forces battalion, and "MCCC immediately," which refers to the Maneuver Captain's Career Course. The bigrams show the unique positions held by each gender in the Army in greater detail than the analysis on individual words. Both inverse document frequency plots reveal that the unique words and phrases that are written about officers are their job titles or duties.

5. Discussion

The findings suggest that male and female officers are objectively rated the same in the United States Army because there is little disparity between the proportion of males and females rated as "Most Qualified," "Highly Qualified," "Qualified," and "Not Qualified." Additionally, after performing sentiment analysis, it was determined that the words most commonly used to describe male and female officers followed similar patterns. The rarer words, when present, described the position or job the officer held, rather than their unique characteristics of their performance. This leads to the conclusion that males and females in the United States Army are both objectively and subjectively rated the same on OERs.

A circumstance that explains the findings of this analysis is the formal structure that OERs tend to follow. Senior raters typically follow the same general format when writing the evaluations for officers, which is why many words have such high frequencies. The frequency analysis shows that the word "potential" appears in nearly 80% of OERs, with over 40% of these instances describing an officer's "unlimited potential." Additionally, raters give the recommendation to either promote with peers, ahead of peers, or behind peers, which may be why the words "promote" and "peers" have a high frequency for both genders.

Although the findings do not show a bias against a specific gender, this reveals that officers oftentimes do not receive candid, honest feedback on their evaluations. The formal, consistent language devalues the narrative if they all are written the same way and creates the question of how much meaning actually goes into the narratives. It is difficult to distinguish between high performers and people who are truly exemplary if most officers are written about using the same language and descriptions. This raises the issue of whether or not assessors actually take time to read narratives, or if they just focus on the block check rating for promotions. If nearly half of the officers are described as having unlimited potential, it is difficult to distinguish what really makes an officer stand out.

6. Limitations and Future Work

The conclusions of this analysis are limited due to the fact that the dataset contains one years worth of OERs which only depicts an officer's performance at a single instance in time. Assumptions were made about the officers without taking into consideration their entire profile, to include successive ratings and block checks. An additional limitation is the fact that there is not much variation in the words used in the senior rater narratives due to the widely accepted standardized format. Gender biases may still exist in the Army, but they might not be apparent in the OER narratives due to the formal language used to write them.

In the future, analysis can be performed on an officer's entire profile that follows them throughout their career to assess how their narratives change over time and correspond to awards, promotions, or jobs they receive. This research can be used to assess whether the words used in the OER narratives for an officer actually have an impact on their career succession, or if the the words do not hold much weight.

7. References

DA Form 67-10-1: Company Grade Plate (O1-O3; WO1-CW2) Officer Evaluation Report. United States Army, November 2015.

Eagly, A.H., & Carli, L.L. (2018). Women and the Labyrinth of Leadership. In Rosenbach, W.E. (Eds.), *Contemporary Issues in Leadership.* New York, NY: Routledge.

Ecklund, M. V. . M. (2006). Leading change: Could a joint OER (officer evaluation reports) be the catalyst of Army transformation? *Military Review*, (1), 71.

Enge, Eric (2015, May 13th). *Inverse Document Frequency and the Importance of Uniqueness.* Retrieved from Moz, Inc website: https://moz.com/blog/inverse-document-frequency-and-the-importance-of-uniqueness

Fallesen, J. (2017). Response to Col. Kevin McAninch's 'how the Army's multi-source assessment and feedback program could become a catalyst for leader development: (Military review, September-October 2016). *Military Review*, (1), 122.

Gupta, V. K., Turban, D. B., Wasti, S. A., & Sikdar, A. (2009). The role of gender stereotypes in perceptions of entrepreneurs and intentions to become an entrepreneur. *Entrepreneurship: Theory and Practice*, (2), 397.

Kite, D.P. (1998). *U.S. Army Officer Evaluation Report; Why Are We Writing to Someone Who Isn't Reading.* (M. A. A. Air Command and Staff Coll., Ed.) (Vol. AU/ACSC/151/1998-04). Non Paid ADAS.

Pang, B. & Lee, L. (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

Parker, K., Horowitz, J. M., & Rohal, M. (2015). Women and Leadership: Public Says Women are Equally Qualified, but Barriers Persist. Washington, DC: Pew Research Center, 1-56.

Silge, J., & Robinson, D. (2019). *Text Mining with R: A Tidy Approach.* O'Reilly.

Smith, D.G., Rosenstein, J.E., Nikolov, M. C., & Chaney, D. A. (2019). The Power of Language: Gender, Status, and Agency in Performance Evaluations. *Sex Roles: A Journal of Research*, (3–4), 159. https://doi.org/10.1007/s11199-018-0923-7

Smith, D.G., Rosenstein, J.E., & Nikolov, M.C. (2018). The Different Words We Use to Describe Male and Female Leaders. *Harvard Business Review*. Retrieved from https://hbr.org/2018/05/the-different-words-we-use-to-describe-male-and-female-leaders

Trobaugh, E. M. (2018). Women, Regardless: Understanding Gender Bias in U.S. Military Integration. *Joint Force Quarterly, 88.*

United States Army Human Resources Command. (2014). *Revised Officer Evaluation Reports.*