Linguistic Analysis of Veteran Job Interviews to Assess Effectiveness in Translating Military Expertise to the Civilian Workforce

Caroline J. Wendt[1], Ehsanul Haque Nirjhar[2], Theodora Chaspari[1]
[1] University of Colorado Boulder, [2] Texas A&M University
{caroline.wendt, theodora.chaspari}@colorado.edu
nirjhar71@tamu.edu

Abstract

The ways in which natural language processing (NLP) can inform how veterans can improve effectiveness in translating military experience to workforce utility is underexplored. We design NLP experiments to evaluate the degree of explanation in veteran job interview responses as a proxy for perceived hireability. We examine linguistic and psycholinguistic features, context, and participant variability to investigate the mechanics of effective communication in employee selection. Results yield good performance when distinguishing between varying degrees of explanation in responses using LIWC features, indicating robustness of linguistic feature integration. Classifying Over- and Under-explained responses reflects challenges of class imbalance and the limitations of tested NLP methods for detecting subtleties in overly verbose or concise communication. Our findings have immediate applications for assistive technologies in job interview settings, and broader implications for enhancing automated communication assessment tools and refining strategies for training and interventions in communication-heavy fields.

1 Introduction

The complexity of verbal communication is a fundamental factor in various realms, including psychology, education, and human-computer interaction (HCI). The degree to which individuals explain themselves reveals insights into their cognitive processes, social interactions, and personality traits. These factors both explicitly and implicitly define the ways in which speakers are perceived, and are thus essential for assessing candidates in structured job interviews (Levashina et al., 2014). The qualifications, background, and training of the majority of military veterans are notably different from job candidates in the general population. Many companies acknowledge that hiring veterans is beneficial, as veterans often posses desirable workforce qualities that arise from their unique experiences, such as strong work ethics, leadership skills, adaptability, team orientation, and professionalism (Sakib et al., 2024). Yet, veterans commonly experience persistent employment challenges post-service due to organizational and societal barriers such as lack of transition support, stressful experiences, and perceived discrimination, as well as personal barriers like incongruence between military and civilian culture (Keeling et al., 2018; Nirjhar et al., 2022). Veterans demonstrate distinct verbal communication gaps in explaining their military experience, references, jargon, and specialized skills relative to the workplace (Mael et al., 2022; Roy et al., 2020; Sakib et al., 2024). Industry interviewers are often unaware of these factors (Mael et al., 2022), further exacerbating the problem with negative stereotypes, stigma, and exclusion (McAllister et al., 2015).

Artificial intelligence (AI) enhances a range of individualized assistive tools to address visual, auditory, cognitive, and physical needs (Zdravkova, 2022). Automated natural language processing (NLP) and understanding can help specific populations communicate and interact with surroundings more effectively and efficiently. One immediate application is intelligent interview training, which provides a suitable environment for individuals to practice and refine relevant verbal and nonverbal behaviors. Such training can help participants adapt to cognitively demanding and socially challenging interview situations (Hemamou et al., 2019a). Given that employment interviews are an immediate obstacle in the hiring process, AI-powered interview training, augmented with NLP, has potential to identify linguistic and communicative behaviors that may hinder candidates' performance, then suggest precise modifications to improve their communication skills (Marienko et al., 2020).

Previous research in intervention technologies for interview training primarily seeks to investigate and improve social skills and positive personality

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 343–355
April 30 - May 1, 2025

signals. Various games, systems, and virtual reality platforms have been developed to help users improve interview performance and stress levels through simulated interactions, providing feedback on behavioral and emotional cues (Anderson et al., 2013; Gebhard et al., 2018; Hoque et al., 2013; Hartholt et al., 2019). Other work has used multimodal data from asynchronous job interviews, analyzing linguistic, acoustic, and visual signals to predict personality traits, hireability, and communication skills, with factors such as word choice, personal pronoun use, and speech fluency shown to significantly impact interview outcomes (Chen et al., 2017; Hemamou et al., 2019a,b; Nguyen and Gatica-Perez, 2016; Muralidhar et al., 2016; Naim et al., 2016).

Departing from prior studies, we present foundational knowledge to improve interview training with several key contributions to enhance the development of intervention technologies that use NLP. While some related studies have contributed to adaptive solutions for specific populations (Hartholt et al., 2019; Marienko et al., 2020), we focus on military veterans, a population encountering distinct difficulties in job interviews. Rather than investigating global characteristics of interviewees, such as personality and overall interview outcomes (Anderson et al., 2013; Gebhard et al., 2018; Hoque et al., 2013; Hartholt et al., 2019), this research provides detailed analysis of turn-level linguistic behaviors that influence verbal communication patterns. We examine dynamic and complex synchronous (instead of static, asynchronous) interactions between interviewers and interviewees. We not only consider interview responses (Verrap et al., 2022), but also account for the content of interview questions, context, turn-taking behaviors, and individualized interviewee variability.

2 Methods

2.1 Data

The data are from a concluded mock job interview study between experienced industry professionals and military veterans in transition to civilian life post-service (Verrap et al., 2022). Interviews were conducted in a hybrid format, where veterans voluntarily participated in the lab, while interviewers joined virtually via Zoom. In total, 38 veterans representing all branches of the military completed the study. The demographic information of participants and interviewers is summarized in Table 1. Participants each received a customized job description created based on their individual qualifications. Participants were thus instructed to act as if they were applying to and interviewing for their unique jobs, and interviewers conducted the calls as they would in their professional roles. Transcript data from the audio and video recordings were automatically generated with Zoom's speech recognition tool, then manually corrected for errors. Response data from the cases in which interviewers asked follow-up questions were aggregated as part of the original question's response.

Three undergraduate psychology students with experience in behavioral coding annotated the interview data (Chorney et al., 2015). The degree of explanation in responses is categorized into four target classes:

- **Under-explained**: Brief and do not fully answer the question, often end abruptly
- **Succinct**: Concise and complete
- **Comprehensive**: Detailed and fully answer the question
- **Over-explained**: Long with excessive detail that can affect coherence

The length (word count) and duration (time in seconds) of responses are correlated ($r(284) = 0.97, p < 0.001$) and tend to increase across these categories. Annotator agreement for the degree of explanation is moderate with Krippendorff's $\alpha = 0.677$, when all samples are included and after adjudication (Krippendorff, 2011). Final labels corresponding to each response were determined by majority voting. Figure 1 shows the imbalanced distribution of the classes at the extremes, with "Under-explained" and "Over-explained" as the minority classes, which are of particular interest due to their negative impact on interview performance and overall perceived hireability.

2.2 Experiments

Rather than pursuing a traditional four-way classification task, we calibrate our experimental approach to the imbalanced nature of the dataset by defining two distinct binary classification problems where we distinguish between (1) Comprehensive and Over-explained responses and (2) Under-explained and Succinct responses. In each of these classification problems, we experiment with NLP feature extraction and selection techniques and optimize performance over various text inputs, representation methods, and linguistic features to gain insight

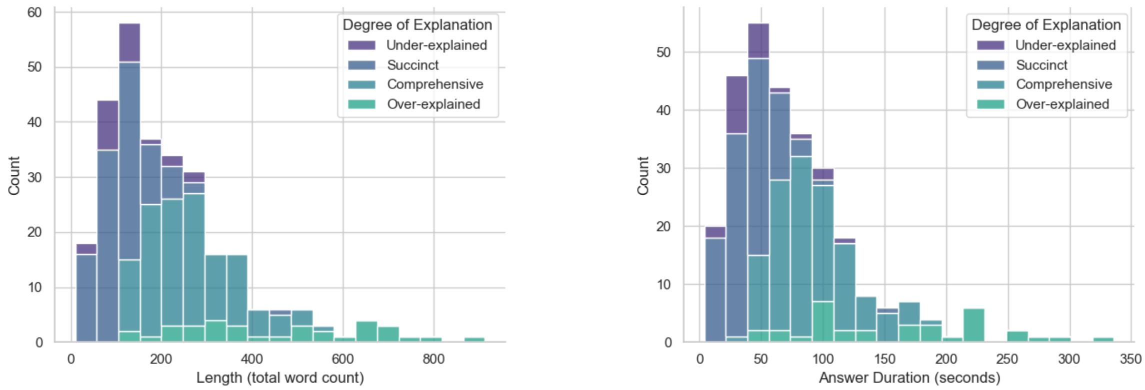

Figure 1: Histograms of total word count and duration of responses per class of degree of explanation. These figures show the dataset's class imbalance, where classes at the extremes are underrepresented.

into what differentiates the level of explanation in veteran responses.

2.3 Features

We use the Linguistic Inquiry and Word Count (LIWC) method to extract a feature set for each input (Boyd et al., 2022). LIWC features are 117 in total and provide a structured and interpretable way to quantify the content of the text by capturing critical aspects of language use, enabling the analysis of linguistic patterns and their relationship to different psychological or social outcomes, which is relevant in the context of job interviews. In our text analysis, for instance, we observe that for LIWC features which capture cognitive processes and perception, Comprehensive responses more frequently contain "causation" language ($t(76.16) = 2.29, p = 0.02$), whereas Over-explained responses more frequently contain "focuspast" language ($t(53.66) = -2.30, p = 0.03$). Causation words (e.g., how, because, make, why) explain why something happened, connecting events or ideas through cause-and-effect relationships, such as when the veteran elaborates on their explanations or justifies their points. Over-explained responses, however, often involve recounting stories or providing excessive context; speakers frequently describe past events, actions, or experiences to justify or elaborate on their point. By contrast, Under-explained responses have a higher frequency of words in the LIWC "tentative" category ($t(52.09) = -2.30, p = 0.03$). These words (e.g., might, could, maybe, not sure) express hesitation and uncertainty, like when the speaker deliberately hedges their statements to avoid being challenged or questioned further, or takes a cautious approach to statements due to low confidence in knowledge or ability to articulate their point or lack of clarity in the question. Political or socially strategic language occurs more frequently in Succinct responses ($t(28.79) = 2.42, p = 0.02$), reflecting topics of governance, politeness markers, and harmonious language. Succinct responses aim to convey necessary information clearly and directly without overloading the interviewer. In doing so, Succinct responses often use language to ensure the response is well-received due to awareness of the interviewer's expectations, while avoiding unnecessary details or uncertain language, and instead focusing on clear and positive expressions.

To capture the syntactic structure of the text and to further analyze patterns in participants' language use, we experiment with 48 part-of-speech (POS) features (Honnibal et al., 2020). For example, we observe that Comprehensive responses tend to include more wh-pronouns (WP) (e.g., (who, what, when, where, why, how) compared to Over-explained responses ($t(84.40) = 2.86, p = 0.01$). Comprehensive responses aim to address key details, provide clarity, and cover the full context of a topic such that this language is often leveraged to introduce or elaborate on specific aspects, answering questions directly and fully. Yet, Over-explained responses tend to contain more personal pronouns (PRP) ($t(57.46) = -2.20, p = 0.03$). A potential reason for this might be that over-explaining often involves recounting personal stories or providing excessive background information, leading to a higher frequency of self-references. Frequent use of personal pronouns tends to overly center the narrative on personal experiences and viewpoints, reflected in Over-explained responses that tend to emphasize the speaker's experiences, actions, and opinions. Succinct responses tend to use more coordinating conjunctions (CC, e.g., and, but, or) because they aim to compactly connect ideas, actions, or clauses within a limited scope ($t(34.15) = 2.12, p = 0.04$). In contrast, Under-explained responses often omit details and connections, resulting in fewer opportunities for conjunctions to bridge ideas effectively. See Table 2 and Table 3.

We reduce each set to the most informative psychological and linguistic data in the text by retaining only the features that are statistically relevant to each classification task. We conduct t-tests to select the POS and LIWC features that significantly differ between classes, where features are considered statistically important for distinguishing between the classes at the 5% significance level.

We additionally experiment with normalized military jargon term counts as a feature for analyzing response texts. Jargon term counts refer to the raw frequency of predefined military-specific phrases (e.g., mission, operation, sergeant) appearing in the text, providing a direct measure of the use of military language (Figure 2). Normalized counts, calculated as the proportion of military terms relative to the total word count of the turn, account for text length, enabling fair relative comparisons of the use of military jargon across responses of varying lengths. These features are explored to test if higher counts may indicate a speaker's familiarity with or connection to military culture, and thus help distinguish between responses.

For text representation, we assess Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) vectorizers, and Bidirectional Encoder Representations from Transformers (BERT) embeddings (Salton et al., 1975; Devlin et al., 2019). The vocabulary sizes after standard NLP preprocessing for the question and response transcript corpora are 1,578 and 3,593, respectively. On average, the dimensions of the BoW and TF-IDF vectors are 3058.32 (range: 1977 – 4056). The BERT embedding dimensionality is 768 across representations.

2.4 Models

Our modeling approach leverages advanced preprocessing, feature extraction, and a robust classification algorithm within a participant-independent evaluation framework. The experimental approach

Figure 2: Word cloud illustrating the frequency of various military jargon terms in the response dataset, where larger font size indicates more frequent.

utilizes machine learning pipelines to preprocess text and extract features for two binary classification tasks (i.e., Under-explained vs. Succinct, and Comprehensive vs. Over-explained). We examine each text representation (i.e., BoW, TF-IDF, BERT) alone for a baseline and in combination with the considered features (i.e., LIWC, POS tags, normalized jargon). These are extracted based on the interviewee's response only, as well as based on the interviewer's question and the interviewee's response.

To control for participant-level variation and maximize the training data available for model fitting, Leave-One-Participant-Out (LOPO) cross-validation is used to evaluate the model. LOPO emulates real-world scenarios where generalization to unseen participants is critical (Figure 3). To further control variability and assess the performance of the features of interest, we use the Extreme Gradient Boosting (XGBoost) classifier across all experiments, configured with the multi-class log loss evaluation metric, 100 trees with a depth of 6, and minimal regularization. We use XGBoost due to its ability to capture complex feature interactions, handle class imbalance, regularize against overfitting, and efficiently scale to diverse, high-dimensional data types such as BERT embeddings and LIWC features. Compared to preliminary experiments with various classifiers (Multinomial Naïve Bayes, Logistic Regression, Linear SVC, Decision Trees, and Random Forests), we find that XGBoost demonstrates both predictive power and robustness within the LOPO evaluation framework.

3 Results

Table 4 and Table 5 provide an evaluation of multiple text classification experiments, comparing the effectiveness of different input configurations, text

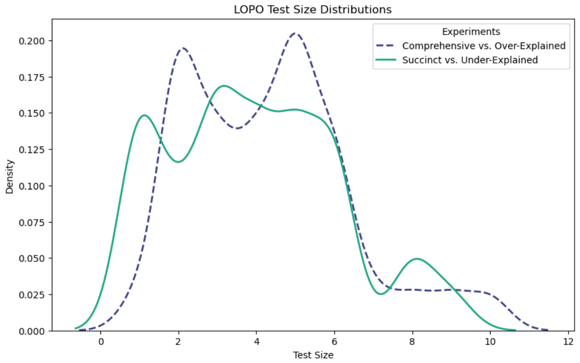

Figure 3: A comparison of the distributions of test sizes between the major experimental categories. The smooth curves represent kernel density estimates, highlighting differences in the spread and concentration of test sizes across experiment types under LOPO cross-validation, where the number of observations associated with each participant varies.

representations, and feature sets. Figure 4 provides an overview of feature performance for the best model results for each feature category across experiments with different inputs. Key insights are summarized below.

In terms of a comparison across features, LIWC features consistently outperform others. Across all setups, the use of LIWC features leads to the best or same overall performance. Over-explained or Under-explained performance (i.e., Class1 F1) also benefit notably from LIWC, suggesting its utility in handling minority or challenging classes. The baseline model, which does not utilize additional features, consistently underperforms compared to models that incorporate LIWC, but tends to perform comparably to other feature sets. Notable gaps are observed in Class1 F1, where the baseline scores range from 0.00 to 0.50, indicating poor detection of the Over-explained and Under-explained responses. However, for the case of distinguishing Under-explained responses, the baseline often performs no worse than more complex models. Models leveraging POS and normalized jargon count features, generally perform similarly to the baseline, with slight improvements in macro F1 and weighted F1 in some cases. For instance, normalized jargon count marginally improves performance over POS in certain cases, but still trails behind the LIWC model performance. Models using both question and response inputs outperform those using only responses in some configurations. Adding question context tends to not improve results significantly for longer responses, but does show some lift when distinguishing between shorter classes, particularly when identifying Under-explained responses. This highlights the importance of leveraging the full conversational context for classification tasks with limited information. For text representation methods, we observe BERT-based representations do not show a clear advantage for these tasks, possibly due to limited feature integration or insufficient fine-tuning. Simpler BoW and TF-IDF representations yield comparable results, but benefit significantly from feature augmentation like LIWC. Performance trends across classes indicate that performance for Succinct and Comprehensive classes, which represent the majority classes, remain high across all setups, with F1 scores consistently above 0.84. This suggests that models can reliably identify less extreme responses regardless of the features used. Over-explained and Under-explained classes remain challenging, with low F1 scores, particularly in baseline and non-LIWC models. This highlights the class imbalance or inherent difficulty in detecting these classes. LIWC consistently improves Over-Explained and Under-Explained F1 scores, e.g., achieving up to 0.50 in classification of Over-explained and 0.21 in Under-explained responses.

4 Limitations and Future Work

A limitation of this study lies in the small data sample. Although difficult to obtain given the interpersonal nature of our dataset, further analyses would benefit from a larger, balanced, and more comprehensively diverse population to improve performance, robustness, and generalizability of algorithms for assistive systems. Increasingly complex data, features, and models, would present greater computational expense. More advanced classification strategies to capture the linguistic subtleties between Comprehensive and Over-explained responses or Succinct and Under-explained responses may possibly require additional data with higher annotator agreement or data augmentation, as well as careful tuning of vectorizers, classifiers, and class weights. Future work could explore advanced integration of LIWC with deep learning approaches, combined feature sets, or fine-tuning BERT embeddings with domain-specific linguistic features to enhance performance. It would be constructive to also investigate the ways in which other linguistic (e.g., reference to military), physical (e.g., body language, posture), and speech (e.g., volume,

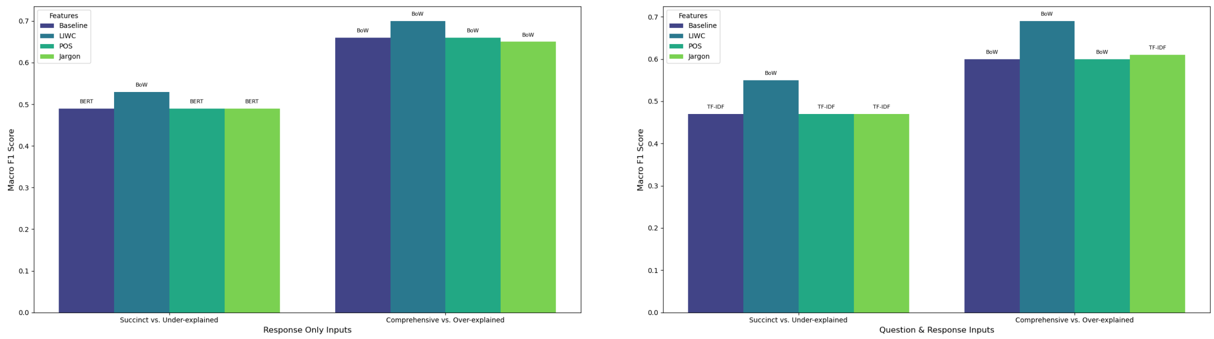

Figure 4: A comparison of feature performance for the best results for each feature category across experiments with response inputs (left) and question and response inputs (right). Bars are colored by feature type and labels above each bar indicate the respective text representation method associated with the best given model. The experiments demonstrate the efficacy of LIWC features for text classification tasks involving nuanced categories like explanation levels. LIWC consistently outperforms baseline and alternative feature sets across all metrics, particularly for the challenging Over- and Under-explained and categories. Combining question and response inputs further boosts model performance, while feature integration remains critical for improving representation-based models like TF-IDF and BERT.

intonation) factors influence the degree of explanation. Future related work should explore these variables in both binary and four-way classification settings. Methods employed and results obtained in our work provide a basis for developing technologies that offer personalized, granular interview feedback in real time. As such, a promising direction for future investigation may involve leveraging large language models and chain-of-thought prompting (Wei et al., 2022) to design interactive interview training interfaces. Specialized applications of further research to narrow communication gaps may extend beyond job interviews to areas like educational assessments and automated dialogue systems. In addition to military veterans, upcoming studies in this space should aim to make interactions more constructive and meaningful for other sensitive groups, such as formerly incarcerated individuals, non-Native speakers, and older adults seeking to re-enter the workforce, by tailoring systems to their unique needs.

5 Conclusion

We use NLP to inform the development of personalized training methods and assistive technologies to aid military veterans in their transition to the civilian workforce. This study integrates advanced linguistic features with robust text representation strategies and participant-dependent cross-validation to detect the degree of explanation in veteran job interview responses. We incorporate LIWC features, which analyze the psychological and cognitive dimensions of text, and POS tagging, which provides syntactic insights, into the text classification pipeline. These features are combined with traditional BoW and TF-IDF vectorization and BERT embedding methods to create a comprehensive feature set that can capture both surface-level and deep linguistic patterns. We advance prior studies by looking beyond the ways in which personal, social, and behavioral impressions and physical characteristics impact interview outcomes (Anderson et al., 2013; Gebhard et al., 2018; Hoque et al., 2013; Hartholt et al., 2019). We also extend existing work by not only considering interview responses, but also accounting for the content of the interview question to understand contextual and turn-taking aspects of conversational communication (Verrap et al., 2022). Classification results from our binary classification experiments reveal that while tested models can generally distinguish between responses with moderate accuracy, correctly identifying certain subclasses within these categories is more challenging, particularly for Under-explained responses. The choice of input features as well as text representation methods significantly impact performance, with LIWC features generally leading to better overall results. This research will contribute to the eventual development of intelligent training technologies that provide personalized learning and reintegration support through mechanisms such as real-time automatic feedback to optimize veterans' job interview outcomes and improve the workforce.

Ethics Statement

Data collection was approved by the institutional review board of the authors' university. All authors strove to maintain highest standards of professional conduct and ethical practice when conducting this work via respecting and maintaining the privacy of participants and security of the data, and disclosing all pertinent system capabilities and limitations.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (#1956021) for this work.

References

Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *International conference on advances in computer entertainment technology*, pages 476–491. Springer.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE.

Jill MacLaren Chorney, C Meghan McMurtry, Christine T Chambers, and Roger Bakeman. 2015. Developing and modifying behavioral coding schemes in pediatric psychology: a practical guide. *Journal of pediatric psychology*, 40(1):154–164.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Gebhard, Tanja Schneeberger, Elisabeth André, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius König, and Markus Langer. 2018. Serious games for training social skills in job interviews. *IEEE Transactions on Games*, 11(4):340–351.

Arno Hartholt, Sharon Mozgai, and Albert" Skip" Rizzo. 2019. Virtual job interviewing practice for high-anxiety populations. In *Proceedings of the 19th ACM international conference on intelligent virtual agents*, pages 238–240.

Léo Hemamou, Ghazi Felhi, Jean-Claude Martin, and Chloé Clavel. 2019a. Slices of attention in asynchronous video job interviews. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019b. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 573–581.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706.

Mary Keeling, Sara Kintzle, and Carl A Castro. 2018. Exploring us veterans' post-service employment experiences. *Military Psychology*, 30(1):63–69.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Julia Levashina, Christopher J Hartwell, Frederick P Morgeson, and Michael A Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1):241–293.

Fred Mael, Will Wyatt, and Uma Janardana Iyer. 2022. Veterans to workplace: Keys to successful transition. *Military Psychology*, 34(5):516–529.

Maiia Marienko, Yulia Nosenko, and Mariya Shyshkina. 2020. Personalization of learning using adaptive technologies and augmented reality. *arXiv preprint arXiv:2011.05802*.

Charn P McAllister, Jeremy D Mackey, Kaylee J Hackney, and Pamela L Perrewé. 2015. From combat to khakis: An exploratory examination of job stress with veterans. *Military Psychology*, 27(2):93–107.

Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. 2016. Training on the job: Behavioral analysis of job interviews in hospitality. In *Proceedings of the 18th acm international conference on multimodal interaction*, pages 84–91.

Iftekhar Naim, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2):191–204.

Laurent Son Nguyen and Daniel Gatica-Perez. 2016. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437.

Ehsanul Haque Nirjhar, Md Nazmus Sakib, Ellen Hagen, Neha Rani, Sharon Lynn Chu, Winfred Arthur, Amir H Behzadan, and Theodora Chaspari. 2022. Investigating the interplay between self-reported and bio-behavioral measures of stress: A pilot study of civilian job interviews with military veterans. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Deborah Roy, Jana Ross, and Cherie Armour. 2020. Making the transition: How finding a good job is a risky business for military veterans in northern ireland. *Military Psychology*, 32(5):428–441.

Md Nazmus Sakib, Ellen Hagen, Nidhal Mazza, Neha Rani, Ehsanul Haque Nirjhar, Sharon L Chu, Theodora Chaspari, Amir H Behzadan, and Winfred Arthur Jr. 2024. Capitalizing on strengths and minimizing weaknesses of veterans in civilian employment interviews: Perceptions of interviewers and veteran interviewees. *Military Psychology*, pages 1–13.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Raghu Verrap, Ehsanul Nirjhar, Ani Nenkova, and Theodora Chaspari. 2022. "am i answering my job interview questions right?": A nlp approach to predict degree of explanation in job interview responses. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 122–129.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Katerina Zdravkova. 2022. The potential of artificial intelligence for assistive technology in education. In *Handbook on Intelligent Techniques in the Educational Process: Vol 1 Recent Advances and Case Studies*, pages 61–85. Springer.

A Appendix

Population	Demographic Feature	Value
Interviewers	*N*	11
	Mean age in years (SD)	44.91 (11.67)
	Male:Female	8:3
	Ethnicity (W, BAA, M)	9, 1, 1
Interviewees (Military Veterans)	*N* completed (total)	38 (41)
	Mean age in years (SD)	40.3 (12.3)
	Male:Female	37:4
	Ethnicity (W, HL, NHPI, A, M, O)	24, 13, 1, 1, 1, 1
	Employed (full, part, not)	25, 4, 12
	Mean years of service (SD)	12.7 (9.1)
	Mean years since end of service (SD)	8.8 (10.6)
	Attended transition assistance	27

Table 1: A summary of the demographic information for the full dataset. The ethnicities represented in the data are abbreviated as follows: White (W), Hispanic or Latino (HL), Black or African American (BAA), Native Hawaiian or Other Pacific Islander (NHPI), Asian (A), Two or More Races (M), and Other (O).

Experiment	Input	Feature	Description	Mean (SD) Class0	Mean (SD) Class1	t-test Result
Comprehensive (0) vs. Over-explained (1)	response	WC	total number of words in the text	262.96 (97.73)	458.56 (206.04)	t(37.23)=-5.37, p<0.01
		BigWords	percentage of words longer than six letters	15.43 (4.64)	12.81 (2.93)	t(84.04)=4.00, p<0.01
		number	percentage of numerical terms (e.g., one, two, 100)	1.17 (1.16)	1.81 (1.29)	t(49.14)=-2.64, p=0.01
		prep	percentage of prepositions (e.g., in, on, about)	13.84 (3.10)	12.72 (2.15)	t(74.65)=2.43, p=0.02
		negate	percentage of negation words (e.g., not, never, no)	1.01 (0.87)	1.51 (1.01)	t(47.54)=-2.59, p=0.01
		Drives	percentage of words related to motivation and needs	5.98 (2.88)	4.93 (2.12)	t(70.53)=2.33, p=0.02
		achieve	percentage of words related to achievement or success	2.05 (1.20)	1.38 (1.00)	t(62.55)=3.35, p<0.01
		Cognition	percentage of words related to thinking and reasoning	14.02 (4.21)	12.54 (3.54)	t(61.50)=2.06, p=0.04
		cogproc	percentage of words related to cognitive processes	12.88 (4.06)	11.10 (3.53)	t(59.55)=2.52, p=0.01
		cause	percentage of words indicating cause and effect	1.82 (1.24)	1.40 (0.85)	t(76.16)=2.29, p=0.02
		tentat	percentage of words expressing uncertainty	3.13 (2.42)	2.22 (1.30)	t(101.24)=2.90, p<0.01
		socbehav	percentage of words related to social actions and interactions	2.90 (1.65)	2.0 (1.08)	t(80.11)=2.95, p<0.01
		work	percentage of words related to working	3.94 (2.45)	2.92 (1.81)	t(70.45)=2.69, p<0.01
		auditory	percentage of words related to hearing or sound	0.22 (0.43)	0.08 (0.20)	t(117.74)=2.80, p<0.01
		focuspast	percentage of words referencing past events	4.30 (2.85)	5.55 (2.80)	t(53.66)=-2.30, p=0.03
		OtherP	percentage of punctuation not categorized as periods, commas, or question marks	2.15 (3.17)	1.09 (2.23)	t(74.41)=2.20, p=0.03
	question	Analytic	a measure of logical and structured thinking based on word patterns	24.18 (23.71)	16.60 (17.91)	t(68.65)=2.02, p=0.04
		conj	percentage of conjunctions (e.g., and, but, or)	7.74 (4.01)	9.54 (4.51)	t(48.48)=-2.11, p=0.04
Succinct (0) vs. Under-explained (1)	response	tentat	see above	2.34 (1.75)	3.40 (2.92)	*t(52.09)=-2.30, p=0.03*
		polite	percentage of words indicating politeness	0.02 (0.08)	0.11 (0.47)	t(125.79)=-1.98 p=0.04
		politic	percentage of words related to political topics	0.78 (0.95)	0.27 (0.78)	t(28.79)=2.42, p=0.02
		health	percentage of words related to health and well-being	0.04 (0.19)	0.22 (0.54)	t(103.24)=-2.80, p<0.01
		illness	percentage of words related to illness or medical conditions	0 (0)	0.06 (0.32)	t(106)=-2.13, p=0.04
		food	percentage of words related to food and eating	0 (0)	0.09 (0.36)	t(106)=-2.55, p=0.01
		auditory	see above	0.02 (0.08)	0.10 (0.38)	t(127.58)=-1.99, p=0.04
		OtherP	see above	1.12 (2.14)	2.61 (3.99)	t(60.39)=-2.53, p=0.01
	question	Authentic	a measure of personal authenticity based on word usage	25.01 (30.61)	43.07 (34.7)	t(29.02)=-2.37, p=0.02
		Tone	a calculated score reflecting positive or negative tone	83.20 (23.94)	69.84 (27.44)	t(28.45)=2.07, p=0.04
		we	percentage of first-person plural pronouns (e.g., we, us, our)	0.43 (1.21)	1.06 (1.89)	t(48.67)=-2.03, p=0.04
		quantity	percentage of words indicating quantity or amount	5.04 (3.73)	3.15 (3.61)	t(31.46)=2.22, p=0.03
		insight	percentage of words reflecting understanding or awareness	2.69 (3.22)	4.61 (4.78)	t(45.67)=-2.36, p=0.02
		tentat	see above	2.76 (2.94)	4.37 (4.74)	t(50.21)=-2.09, p=0.04
		emo_neg	percentage of words expressing negative emotions	0 (0)	0.20 (1.02)	t(106)=-2.04, p=0.04
		tech	percentage of words related to technology	0.03 (0.14)	0.27 (0.83)	t(125.92)=-2.76, p<0.01
		want	percentage of words expressing desire	0.04 (0.18)	0.22 (0.70)	t(123.56)=-2.38, p=0.02
		Perception	percentage of words related to perception (e.g., look, feel).	3.63 (3.54)	6.52 (4.72)	t(40.71)=-3.33, p<0.01
		attention	percentage of words indicating focus or attention	0.14 (0.46)	0.51 (1.14)	t(87.60)=-2.56, p=0.01
		motion	percentage of words related to movement	0.60 (0.91)	1.14 (1.68)	t(59.33)=-2.19, p=0.03
		space	percentage of words related to space and location	2.36 (2.92)	3.93 (3.34)	t(35.50)=-2.28, p=0.03
		time	percentage of words related to time	1.32 (1.88)	2.75 (3.24)	t(54.65)=-2.85, p<0.01
		OtherP	see above	1.71 (3.16)	3.37 (4.83)	t(47.15)=-2.06, p=0.04

Table 2: Significant LIWC feature t-test results for the various experiments. We use an independent samples t-test. The t-statistic indicates how much the means of the two groups differ relative to the variation in the sample data. We consider $p < 0.05$ to be statistically significant, meaning there is strong evidence against the null hypothesis of no difference between the groups, such that the observed difference in means is unlikely to have occurred by random chance. Here, we do not assume equal variance, utilizing Welch's t-test. As an interpretation example, suppose we are comparing the LIWC scores for the word count feature, where Class0 indicates Comprehensive responses and Class1 indicates Over-explained responses. A negative t-statistic would imply that the average word count of Comprehensive responses is lower than that of Over-explained responses. The small p-value in this case supports the conclusion that the long responses statistically tend to have more words compared to the short responses.

Experiment	Input	Feature	Description	Mean (SD) Class0	Mean (SD) Class1	t-test Result
Comprehensive (0) vs. Over-explained (1)	response	PRP	personal pronoun (e.g., I, you, he, she, it, we, they)	0.11 (0.03)	0.13 (0.03)	t(57.46)=-2.20, p=0.03
		VBZ	verb, 3rd person singular present (e.g., runs, talks, is)	0.03 (0.02)	0.03 (0.01)	t(68.74)=2.01, p=0.04
		CD	cardinal number (e.g., one, two, 3, 100)	0.01 (0.01)	0.02 (0.01)	t(45.77)=-2.20, p=0.03
		VBD	verb, past tense (e.g., ran, talked, was)	0.03 (0.03)	0.05 (0.03)	t(55.14)=-2.50, p=0.02
		VBG	verb, gerund or present participle (e.g., running, talking)	0.03 (0.01)	0.02 (0.01)	t(61.51)=2.17, p=0.03
		HYPH	hyphen	<0.01 (0.01)	<0.01 (<0.01)	t(88.35)=2.43, p=0.02
		WP	wh-pronoun (e.g., who, what, whom, which)	0.01 (0.01)	0.01 (<0.01)	t(84.40)=2.86, p=0.01
	question	RB	adverb (e.g., quickly, silently, very, too)	0.06 (0.04)	0.08 (0.04)	t(52.47)=-2.22, p=0.03
Succinct (0) vs. Under-explained (1)	response	CC	coordinating conjunction (e.g., and, or, but, yet)	0.05 (0.02)	0.04 (0.02)	t(34.15)=2.12, p=0.04
	question	VBP	verb, non-3rd person singular present (e.g., run, talk, are)	0.05 (0.03)	0.07 (0.04)	t(40.61)=-3.54, p<0.01
		NNS	plural noun (e.g., dogs, cars, ideas)	0.01 (0.01)	0.03 (0.03)	t(61.75)=-3.76, p<0.01
		POS	possessive ending ('s)	0 (0)	<0.01 (<0.01)	t(106)=-2.06, p=0.04

Table 3: Significant POS feature t-test results for the various experiments. We use an independent samples t-test. The t-statistic indicates how much the means of the two groups differ relative to the variation in the sample data. We consider $p < 0.05$ to be statistically significant, meaning there is strong evidence against the null hypothesis of no difference between the groups, such that the observed difference in means is unlikely to have occurred by random chance. Here, we do not assume equal variance, utilizing Welch's t-test. See the interpretation example in Table 4.

Experiment	Input	Text Representation	Features	Class0 F1	Class1 F1	Macro F1	Weighted F1
Comprehensive (0) vs. Over-explained (1)	response	BoW	none (baseline)	0.87	0.44	0.66	0.78
			LIWC	**0.88**	**0.48**	**0.68**	**0.79**
			POS	0.87	0.44	0.66	0.78
			jargon count	0.86	0.43	0.64	0.77
			normalized jargon count	0.86	0.44	0.65	0.77
		TF-IDF	none (baseline)	0.86	0.18	0.52	0.71
			LIWC	**0.89**	**0.41**	**0.65**	**0.78**
			POS	0.86	0.18	0.52	0.71
			jargon count	0.86	0.22	0.54	0.72
			normalized jargon count	0.87	0.26	0.56	0.73
		BERT	none (baseline)	0.86	0.09	0.47	0.69
			LIWC	**0.90**	**0.45**	**0.67**	**0.80**
			POS	0.86	0.09	0.47	0.69
			jargon count	0.86	0.09	0.47	0.69
			normalized jargon count	0.86	0.09	0.47	0.69
	question & response	BoW	none (baseline)	0.85	0.35	0.60	0.75
			LIWC	**0.90**	**0.54**	**0.72**	**0.82**
			POS	0.85	0.35	0.60	0.75
			jargon count	0.86	0.38	0.62	0.75
			normalized jargon count	0.84	0.35	0.59	0.73
		TF-IDF	none (baseline)	0.87	0.26	0.56	0.73
			LIWC	**0.89**	**0.48**	**0.69**	**0.80**
			POS	0.87	0.26	0.56	0.73
			jargon count	0.87	0.26	0.56	0.73
			normalized jargon count	0.88	0.33	0.61	0.76
		BERT	none (baseline)	0.86	0.05	0.45	0.68
			LIWC	**0.88**	**0.28**	**0.58**	**0.75**
			POS	0.86	0.05	0.45	0.68
			jargon count	0.86	0.05	0.45	0.68
			normalized jargon count	0.86	0.05	0.45	0.68

Table 4: Classification results for the Comprehensive vs. Over-explained experiments with specified text representation methods and features. "Class0" or "Class1" refers to the class listed first or second in the "Experiment." Bold text indicates the best model performance for each experiment.

Experiment	Input	Text Representation	Features	Class0 F1	Class1 F1	Macro F1	Weighted F1
Succinct (0) vs. Under-explained (1)	response	BoW	none (baseline)	0.87	0.06	0.47	0.73
			LIWC	**0.89**	**0.19**	**0.54**	**0.76**
			POS	0.87	0.06	0.47	0.73
			jargon count	0.87	0.06	0.47	0.73
			normalized jargon count	0.87	0.06	0.47	0.73
		TF-IDF	**none (baseline)**	**0.89**	**0.00**	**0.44**	**0.73**
			LIWC	0.88	0.00	0.44	0.72
			POS	0.89	0.00	0.44	0.73
			jargon count	0.89	0.00	0.44	0.73
			normalized jargon count	0.89	0.00	0.44	0.73
		BERT	**none (baseline)**	**0.90**	**0.08**	**0.49**	**0.75**
			LIWC	0.90	0.08	0.49	0.75
			POS	0.90	0.08	0.49	0.75
			jargon count	0.90	0.08	0.49	0.75
			normalized jargon count	0.90	0.08	0.49	0.75
	question & response	BoW	none (baseline)	0.89	0.00	0.44	0.73
			LIWC	**0.90**	**0.26**	**0.58**	**0.79**
			POS	0.89	0.00	0.44	0.73
			jargon count	0.89	0.00	0.44	0.73
			normalized jargon count	0.89	0.00	0.44	0.73
		TF-IDF	none (baseline)	0.88	0.07	0.47	0.73
			LIWC	**0.89**	**0.14**	**0.51**	**0.76**
			POS	0.88	0.07	0.47	0.73
			jargon count	0.88	0.07	0.47	0.73
			normalized jargon count	0.88	0.07	0.47	0.73
		BERT	none (baseline)	0.90	0.00	0.45	0.74
			LIWC	**0.89**	**0.07**	**0.48**	**0.75**
			POS	0.90	0.00	0.45	0.74
			jargon count	0.90	0.00	0.45	0.74
			normalized jargon count	0.90	0.00	0.45	0.74

Table 5: Classification results for the Succinct vs. Under-explained experiments with specified text representation methods and features. "Class0" or "Class1" refers to the class listed first or second in the "Experiment." Bold text indicates the best model performance for each experiment.