“Am I Answering My Job Interview Questions Right?”: A NLP Approach to Detecting the Degree of Explanation in Job Interview Responses

Raghu D. Verrap
Texas A&M University
raghudv@tamu.edu

Ehsanul Haque Nirjhar
Texas A&M University
nijrhar71@tamu.edu

Ani Nenkova
Adobe Research
nenkova@adobe.com

Theodora Chaspari
Texas A&M University
chaspari@tamu.edu

Abstract

Providing the right amount of explanation in an employment interview can help the interviewee effectively communicate their skills and experience to the interviewer and convince that she/he is the right candidate for the job. This paper examines natural language processing (NLP) approaches, including word-based tokenization, lexicon-based representations, and pre-trained embeddings with deep learning models, for detecting the degree of explanation in a job interview response. These are exemplified in a study of 24 military veterans who are the focal group of this study, since they can experience unique challenges in job interviews due to the unique verbal communication style that is prevalent in the military. Military veterans participated in mock interviews with industry recruiters and data from these interviews were transcribed and analyzed. Results indicate that the feasibility of automated NLP methods for detecting the degree of explanation in an interview response. Features based on tokenizer analysis are the most effective in detecting under-explained responses (i.e., 0.29 F1-score), while lexicon-based methods depict the higher performance in detecting over-explanation (i.e., 0.51 F1-score). Findings from this work lay the foundation for the design of intelligent assistive technologies that can provide personalized learning pathways to job candidates, especially those belonging to sensitive or under-represented populations, and helping them succeed in employment job interviews, ultimately contributing to an inclusive workforce.

1 Introduction

Artificial intelligence (AI) can empower a plethora of assistive tools for enhancing one’s visual, hearing, communication, cognitive, and motor skills (Zdravkova, 2022). By automating natural language processing (NLP) and understanding, AI technologies can enable individuals who belong to sensitive populations, to better express themselves or better understand the world around them. Intelligent interview training is one such technology that can facilitate training in a safe environment on specific verbal and nonverbal behaviors and can help individuals effectively adapt to cognitively demanding and socially challenging interview situations (Hemamou et al., 2019b). This technology can further contribute to an inclusive workforce. Since the employment interview comprises the first step of the job hiring process, intelligent interview training augmented with NLP can detect linguistic and semantic communicative behaviors that might jeopardize candidates’ performance in the interview, suggest the exact modifications needed to effectively communicate their skills, and facilitate access to training material and information in a personalized manner (Marienko et al., 2020).

Military veterans is a group that can particularly benefit from assistive interview training technologies. In many countries around the world, military veterans face major barriers to participating in the civilian workforce after separation from active duty (McAllister et al., 2015; Ahern et al., 2015). The military background and training of most veterans is significantly different compared to the general job candidate population, who usually comprise of relatively younger fresh college graduates. Military veterans often find it challenging to clearly articulate their strengths and “brag" about their achievements in the civilian employment interview setting. Particularly, they can experience unique verbal communication gaps, such as ineffective translation of relevant military experience and technical skills, over-explaining their responses, and excessive use of military jargon, that hamper them from successfully obtaining a job in the civilian workforce (Roy et al., 2020). Intelligent job interview training systems can potentially track these linguistic behaviors of interest and provide military veterans the right feedback at the right time.

We conduct a linguistic analysis of veterans’ responses in civilian interview settings. We focus on the degree of explanation in the response, since

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 122 - 129
December 7, 2022 ©2022 Association for Computational Linguistics

this construct is particularly relevant to the interview success and unexplored by previous work, and particularly we investigate a range of NLP systems to detect over/under-explained, succinct, and comprehensive responses (Hagen et al., 2022). To accomplish this task, we examine NLP systems that rely on text tokenization, lexicon-based analysis, and deep learning methods. These are evaluated on transcripts from mock interviews between 24 military veterans and 5 industry recruiters. A total of 163 responses provided during the interviews were coded by third-party annotators with respect to the degree of explanation. Results indicate the feasibility of automated NLP analysis for detecting the outcome of interest. Particularly, features based on tokenizer analysis are the most effective in detecting under-explained responses (i.e., 0.29 F1-score), while lexicon-based methods depict the higher performance in detecting over-explanation (i.e., 0.51 F1-score). Challenges that were met during data analysis, namely, the small data sample, subjectivity in coding, and uneven class distributions, are described. Discussion of these results further provides ways in which the proposed NLP analysis can contribute to the design of assistive technologies for interview training.

2 Related Work

Prior work in assistive technologies for interview training has focused on helping users demonstrate effective social skills and positive personality cues. The TARDIS project, for example, designed a game simulation platform through which interviewees interacted with a virtual agent in an effort to improve social cues and affective expressions during the interview (Anderson et al., 2013; Gebhard et al., 2018). The system automatically detected and analyzed smiles, head nods, and body movements, which were used by a machine learning algorithm to classify the mental state (e.g., stressed, bored, hesitant) and affective state (e.g., positive/negative mood) of the user. During the virtual interview, the user received credits in the game when depicting behaviors that were deemed as effective for the interview. At the end, users received a series of statistics for each of the focal behaviors, which were also visualized over time. MACH—My Automated Conversation coacH is another automated interview training system that provided feedback to the user regarding their performance based on the analysis of facial expressions, speech, and prosody (Hoque et al., 2013). Similarly, Hartholt *et al.* designed a virtual reality system that simulated various interview settings, including the interviewer's propensity toward the interviewee (i.e., friendly, neutral, unfriendly) and the physical space of the interview (e.g., break room, office) (Hartholt et al., 2019). A user would interact with the training system by starting from easy to more challenging scenarios. No additional feedback was provided to the user.

Another line of work has evaluated interviewees based on multimodal data that were mostly collected in an asynchronous manner. Chen *et al.* estimated applicants' personality traits based on the audiovisual analysis of monologue job interviews (Chen et al., 2017). Linguistic analysis was conducted with a Bag-Of-Words text representation. Hemamou *et al.* designed a hierarchical attention model, called "HireNet" that predicted the hirability of an interviewee based on asynchronous video interviewing. HireNet relied on multimodal information from text, audio, and video (Hemamou et al., 2019a,b). Similarly, Ngugen & Gatricia-Perez and Muralidhar *et al.* analyzed acoustic and visual cues of video resumes and examined their effectiveness in estimating the candidate's hireability and social and communication skills (Nguyen and Gatica-Perez, 2016; Muralidhar et al., 2016). Finally, Naim *et al.* analyzed interviewees' performance in mock job interviews using their facial expressions (e.g., smiles, head gestures, facial tracking points), language (e.g., word counts, topic modeling), and prosodic information (e.g., pitch, intonation, and pauses). Results presented in the MIT Interview Dataset suggest that the use of unique words and personal pronouns, and the degree of speech fluency significantly affect one's interview performance (Naim et al., 2016).

The contributions of this paper in comparison to prior work are: (1) While previous work focuses on global characteristics of the interviewee (e.g., personality, social/communication skills) and overall descriptors of the interview outcome (e.g., hireability, performance), this paper provides a closer study to turn-level behaviors that can affect the job interview outcome, thus laying out the foundation toward intelligent assistive technologies that can analyze micro-level data and provide users with detailed feedback at the turn-level; (2) In contrast to the majority of prior work, this paper analyzes data from synchronous interactions between an interviewer and an interviewee, which are more dynamic and diverse; and (3) Prior work has mostly focused on college students or fresh college graduates, while this research investigates a unique population that comprises of military veterans facing unique challenges when preparing for a job interview, thus outlining unique design characteristics when it comes to creating assistive technologies for this population.

3 Data

3.1 Data Collection

We use data from an ongoing research study with U.S. military veterans who participated in a mock job interview conducted by experienced interviewers from the industry. Currently, 24 participants completed the study. Data from one participant is excluded from this paper due to technical issues in pre-processing. The average age of participants was 36.4 years (stand. dev. = 10.6 years), and two out of the 24 participants were female. The study was conducted in a hybrid format, where the interviewees (i.e., military veterans) were present in the lab, and the interviewers (i.e., industry experts) were connected via Zoom video conferencing. In order to obtain naturalistic conversational data in the mock job interview, we created customized job postings tailored to each participant's résumé, which were shared with both the interviewees and the interviewers. Interviewees were instructed to think that they applied for the aforementioned job and they were participating in the corresponding job interview. The interviewers were instructed to conduct the interview based on the job posting, and ask questions in a similar fashion as they would normally do as part of their job role. The average length of the interviews was about 18 minutes (stand. dev. = 6.4 minutes). Audio and video of the interviews were recorded, while the transcripts of the interviews were obtained by the automatic speech recognition functionality provided via Zoom. Transcripts were manually checked for errors, such as spelling mistakes, incomprehensible words, disfluencies, and non-verbal vocalizations. Next, interviews were checked manually to mark the start and end timestamps of each question and their corresponding responses. If the interviewer provided any prompts or asked for additional information after a response, these turns were considered as a part of the response to the original question. In total, 163 responses to the interview questions from the participants were recorded and were used for further analysis. This study has been approved by the institutional review board of the authors' university.

Degree of Explanation	No. of Samples
Under-explained	16
Succinct	67
Comprehensive	58
Over-explained	17
Total Samples	158

Table 1: Distribution of classes characterizing the degree of explanation to an interview question.

3.2 Behavioral Annotation

In order to label the degree of explanation in the responses to the interview questions, behavioral annotation was performed by three third-party annotators, who were undergraduate students in psychology and had previous experience in behavioral coding and annotation tasks. Consistently with previous work (Busso et al., 2016; Lefter et al., 2014), annotators were asked to watch the individual questions and the corresponding responses from the interview and rate the degree of explanation in each response into the following four possible categories. **Under-explained (Class 0)**: Short response that does not fully answer the interviewer's question. Such responses might end abruptly; **Succinct (Class 1)**: Concise and to-the-point responses that answer the interviewer's question fully and briefly; **Comprehensive (Class 2)**: Detailed response that answers the fully answers the question; and **Over-explained (Class 3)**: Very long response to the question with excess verbiage and too much detail that potentially affects the coherence of the answer.

The numerical labels are assigned based on the expected increasing order in response length for each of these categories (i.e., succinct responses are expected to be shorter compared to comprehensive ones). The annotation process resulted in a moderate annotator agreement of Fleiss' $\kappa = 0.437$ (Fleiss, 1971; Hallgren, 2012). After the annotation, five responses yielded labels with complete disagreement. These were excluded from the rest of the analysis, which renders the sample size, $N = 158$. The final labels were obtained by aggregating annotations through majority voting. Table 1 shows the distribution of labels obtained from this aggregation. It is to be noted that both "Under-explained" and "Over-explained" classes are minority classes, although they are the classes of interest, since these types of responses tend to contribute most to perceived hireability and job interview performance.

4 Methods

Since the numbers of samples belonging to the classes of interest (i.e., "Under-explained", "Over-explained") is much lower compared to the majority classes, it would be counter-productive to formulate the target problem as a 4-way classification task. To resolve this issue, we examine the association between the response length and the explanation labels. Intuitively, we anticipate that responses belonging to the "Under-Explained" and "Succinct" classes will have significantly shorter length compared to the ones belonging to the "Comprehensive" and "Over-Explained" classes. Response length is measured in terms of word count (i.e., the number of words in the response) and response duration (i.e., the duration of the response in seconds). Both these measures exhibit significantly high Pearson's correlation coefficients with the explanation labels (i.e., $r = 0.68, p < 0.01$ for word count, $r = 0.66, p < 0.01$ for response duration). This suggests that the shorter responses tend to fall into "Under-explained" and "Succinct" categories, while the longer responses belong to the "Comprehensive" and "Over-explained" classes. To further confirm this, a binary classification task is conducted to identify whether a response falls into the short (i.e., "Under-explained", "Succinct") or long (i.e., "Comprehensive", "Over-explained") category. For this purpose, a logistic regression model with response length as feature and with leave-one-subject-out cross-validation is used, which resulted in an macro-average F1-score of 0.87. This suggests that we can simply classify the responses into the short (i.e., "Under-explained", "Succinct") or long (i.e., "Comprehensive", "Over-explained") category before estimating the original classes. Therefore, to estimate the degree of explanation, in the following analysis, we formulate two binary classification problems (i.e., "Under-explained" vs. "Succinct", "Comprehensive" vs. "Over-explained") instead of a 4-class problem.

We pursue three different approaches for these binary classification tasks. The first approach employs a tokenizer that breaks text into word tokens, followed by a decision tree that conducts the binary classification task. The second approach utilizes a lexicon-based model of psycholinguistic speech attributes, followed by a decision tree. The third approach leverages a transformer-based model pre-trained on a large corpus of English text in self-supervised manner. Since the classes of each of the binary classification tasks are unbalanced, the F1-score is used as evaluation metric for the following systems. F1-score is reported for each class using a leave-one-subject-out cross-validation. According to this, the responses from one interviewee are included in the test set and the responses from the remaining interviewees are included in the train set, with this procedure repeating until all interviewees are part of the test set.

4.1 Tokenizer

We extract the linguistic information from the participants' responses to the interview questions using NLTK tokenizer (Bird et al., 2009). The NLTK tokenizer breaks each response into chunks at the word-level that can be considered as discrete elements. Tokens are generated from the response text without any truncation and padding. A total of 510 tokens with frequency more than three are selected as features for conventional machine learning models. The frequency of the corresponding tokens serves as the feature vector of length 510 to a decision tree model that conducts the binary classification tasks.

4.2 Lexicon-based method

In order to identify the psycholinguistic content of the participants' responses to the interview questions, we employ the Linguistic Inquiry and Word Count (LIWC) toolbox (Pennebaker et al., 2015). This tool measures the count (or percentage) of words from several constructs, known as LIWC categories. The LIWC categories include general descriptors (e.g., word count, words per sentence), summary variables (e.g., analytical thinking, clout), standard linguistic dimensions (e.g., pronouns, verbs), psychological constructs (e.g., affect, cognition), personal concern constructs (e.g., work, leisure), informal language marker (e.g., filler words, assents), and punctuation (e.g., periods, commas). Overall, we obtain 93 LIWC features from each sample, that comprise the input features of a binary decision tree.

4.3 Deep learning method

We further explore the use of deep learning models for the considered binary classification tasks. We use the RoBERTa-base (Liu et al., 2019) as the backbone network, a popular transformer-based model (Vaswani et al., 2017) pre-trained on a large corpus of English text in self-supervised manner. The input of this model comprises of the segments resulting from the Tokenizer (Section 4.1), namely, the first 510 tokens. The input is connected to two fully connected layers with 768 nodes each, ReLU activation, and dropout, following by the final output layer. As the dataset is highly unbalanced, we perform undersampling on the majority class and oversampling on the minority class. In addition, we freeze the initial 75% layers of the RoBERTa base pre-trained model. The model is trained for 20 epochs with a learning rate of 10^{-5}.

5 Experiments

Results obtained by the different NLP systems are summarized in Table 2. The F1-score for the "Succinct" and "Comprehensive" classes is significantly higher than the other two, since these are the majority classes. The deep learning method that relies on the RoBERTa model further achieves higher score than the Tokenizer and Lexicon-based methods for the "Succinct" and "Comprehensive" classes. This is anticipated as these two classes have a relatively high number of samples, thus the deep learning model can effectively learn their linguistic representation. Meanwhile, the lexicon-based features achieve the highest performance for the "Over-explained" class, which might be due to the fact that these two types of responses can be effectively differentiated via psycholinguistic dimensions. Statistical analysis via t-tests between the two classes of interest indicates that comprehensive responses depict significantly more positive emotional tone compared to over-explained responses ($\mu_3 = 56.83\%$, $\mu_4 = 44.47\%$, $p < 0.05$), where μ_3 and μ_4 are the mean values of the comprehensive and over-explained responses, respectively. This might be attributed to the fact that over-explained responses merely report content without depicting one's affective view. Comprehensive responses also include a significantly larger percentage of long words (i.e., words greater than six letters) compared to over-explained responses ($\mu_3 = 17.14\%$, $\mu_4 = 13.67\%$, $p < 0.01$) and significantly more work-relevant words ($\mu_3 = 4.88\%$, $\mu_4 = 3.39\%$, $p < 0.05$). This indicates that comprehensive responses are characterized by more complex expression (Smith-Keiling and Hyun, 2019) and communicate one's work-related experiences. On the contrary, over-explained responses have a significantly larger number of male references compared to comprehensive ones ($\mu_3 = 27.24\%$, $\mu_4 = 67.55\%$, $p < 0.05$) and include more past tense verbs ($\mu_3 = 3.87\%$, $\mu_4 = 5.39\%$, $p < 0.05$), potentially because over-explained responses are overly focused on one's immersion to past military experiences which are typically associated with male references. Finally, the Tokenizer method achieves the highest F1-score for the "Under-explained" class, potentially because these types of responses depict distinctive patterns with respect to the frequency of tokens compared to the "Succinct" class.

6 Discussion

The increasingly complex and demanding employment market and future workforce requires mature handling of content and emotions by the job candidates, therefore failing to explain one's skills or over-sharing information can be detrimental to succeeding in the employment interview (Cismas, 2021). Results from this study indicate that various types of NLP techniques can be effective in automatically identifying the degree of explanation in job interview responses, which can be particularly valuable when designing training technologies to prepare candidates for future employment. While previous work has focused on behavioral impressions that can affect the overall outcome of the interview (Anderson et al., 2013; Gebhard et al., 2018; Hoque et al., 2013; Hartholt et al., 2019), this paper focuses on linguistic behaviors at the turn-level, which can serve as the foundation for providing tangible low-level feedback to the interviewee. Training technologies that rely on automated NLP systems, such as the ones examined in this paper, can help pinpoint exact turns in the dialog that effectively serve the job interview outcome (i.e., succinct, comprehensive responses), as well as turns that might hurt the interview outcome (i.e., under-explaining, over-explaining). Intelligent cognitive enhancement technologies can potentially assist job candidates in helping them effectively communicate their skills to the interviewers. Such technologies need to rely on robust NLP approaches, that are adequately generalizable to unseen users and new contexts and depict reliable performance, especially for the detection of classes of interest, such as the under-explaining and over-explaining classes in our case. In addition, NLP technologies need to be effectively meshed with human-computer interaction (HCI) interfaces, in order to provide feedback in the right form (e.g., visual, tactile) and the right time (e.g., during practice, post-practice). In addition to detecting points of improvement, explaining their role in interview performance and suggesting appropriate changes to those responses would pave the way for personalized learning pathways. It is also essential to consider the degree of expla-

Methods	F1-score			
	Under-explained	Succinct	Comprehensive	Over-explained
Tokenizer	0.29	0.81	0.74	0.26
Lexicon-based method	0.22	0.78	0.83	0.51
Deep learning method	0.27	0.89	0.84	0.39

Table 2: F1-score for each class of interest obtained by the considered methods.

nation in the context of other linguistic behaviors (e.g., excessive use of military jargon, ineffective translation of military experience to the civilian job context), gestures (e.g., rigidity in posture), and vocal expressions (e.g., voice loudness), which will allow us to design technologies that can assist veteran interviewees in a holistic manner. User studies are needed to be conducted so that we can better understand the effectiveness of these technologies in the overarching goal of assisting military veterans to succeed in civilian job interviews.

7 Conclusion

We examined linguistic behaviors of military veterans that are indicative of the degree of explanation in job interview responses. We investigated different types of linguistic descriptors, ranging from word-based tokenization and lexicon-based representations, to pre-trained embeddings with deep learning models. Our results indicate that pretrained embeddings are effective in detecting succinct and comprehensive responses, which contain the majority of samples. Lexicon-based features can reliably detect over-explained responses, potentially because of their unique psycholinguistic characteristics related to affect, work experience, and complex expression. Finally, under-explained answers are best recognized via the token-based approach, which might be due to the fact that these are characterized by significantly different frequency of tokens compared to the succinct responses. Results from this study lay the foundation toward intelligent interview training technologies that provide personalized learning by detecting verbal behaviors important for the job interview, explaining their role to the user, and suggesting appropriate changes that can effectively help users secure their desired job.

Limitations

The results of this work should be considered in the light of the following limitations. First, while it is difficult to obtain large-scale corpora from real-life interpersonal interactions, the relatively small size of the dataset prevents results of this study from adequately generalizing to other individuals and populations. In addition, due to the demographics of the region from which the data was sampled, the current dataset is highly skewed toward White male participants. As part of our future work, we will be verifying those findings with additional data that will include more diverse participants, which will allow us to make these technologies truly inclusive to all people. Second, the moderate agreement level (i.e., $\kappa = 0.437$) will be addressed via adjudication meetings. Third, this work takes into account the interviewee's response in isolation without considering the content of the question. Future work will incorporate the interview context, turn-taking between interviewer and interviewee, and acoustic information from speech, which is expected to yield improved performance.

Ethics Statement

The authors of this paper strove to maintain highest standards of professional conduct and ethical practice when conducting this work via respecting and maintaining the privacy of the participants of this study and security of the data and disclosing all pertinent system capabilities and limitations. This work is guided by the values of equality, inclusiveness, and respect for others, since it aims to render assistive interview technologies accessible to populations such as military veterans who have traditionally faced challenges in entering the workforce and have not actively been the focus of prior studies in computing that have examined the automated processing of interview data.

Acknowledgements

The authors would like to acknowledge the National Science Foundation (NSF) for funding this work (NSF #1956021, NSF #1955721, the Don & Ellie Knauss Veteran Resource and Support Center (VRSC) at Texas A&M University for their partnership in this research, and undergraduate students, Luis Garcia, Li Wen Jan, Albin Kyle Myscich, and Felicity Woodson for their contribution in data transcription, annotation, and pre-processing.

References

Jennifer Ahern, Miranda Worthen, Jackson Masters, Sheri A Lippman, Emily J Ozer, and Rudolf Moos. 2015. The challenges of afghanistan and iraq veterans' transition from military to civilian life and approaches to reconnection. *PloS one*, 10(7):e0128599.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

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.

Suzana Carmen Cismas. 2021. Strategies to enhance students' employability and job interview abilities by didactic role-plays. *Reading Multiculturalism. Human and Social Perspectives*, page 23.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

Ellen Hagen, Md Nazmus Sakib, Neha Rani, Ehsanul Haque Nirjhar, Ani Nenkova, Theodora Chaspari, Sharon Lynn Chu, Amir Behzadan, and Winfred Arthur, Jr. 2022. Interviewer perceptions of veterans in civilian employment interviews and suggested interventions. International Military Testing Association.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

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.

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.

Iulia Lefter, Gertjan J Burghouts, and Leon JM Rothkrantz. 2014. An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces*, 8(1):29–41.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

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.

Beverly L Smith-Keiling and Hye In F Hyun. 2019. Applying a computer-assisted tool for semantic analysis of writing: Uses for stem and ell. *Journal of microbiology & biology education*, 20(1):70.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Katerina Zdravkova. 2022. The potential of artificial intelligence for assistive technology in education. In *Handbook on Intelligent Techniques in the Educational Process*, pages 61–85. Springer.