A LINGUISTIC ANALYSIS TO QUANTIFY OVER-EXPLANATION AND UNDER-EXPLANATION IN JOB INTERVIEWS

An Undergraduate Research Scholars Thesis

by

ALBIN KYLE MYSCICH

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Theodora Chaspari

May 2022

Major: Computer Science

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Albin Kyle Myscich, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisor, Dr. Theodora Chaspari, prior to the collection of any data used in this final thesis submission.

This project required approval from the Texas A&M University Research Compliance & Biosafety office.

TAMU IRB #: 2020-0709 Approval Date: 04/13/2021

TABLE OF CONTENTS

Page

ABSTRACT

A Linguistic Analysis to Quantify Over-Explanation and Under-Explanation in Job Interviews

Albin Kyle Myscich
Department of Computer Science
Texas A&M University

Research Faculty Advisor: Dr. Theodora Chaspari
Department of Computer Science
Texas A&M University

Receiving insight into the thoughts and feelings of a recruiter is vital to understanding effective job interviews. To ascertain categorical responses and speech patterns, audio and visual data from mock job interviews were collected between interviewees and company representatives. From the study, extracted features of audio and visual data were compiled. As a result, several approaches involving deep learning were leveraged to infer the probability of an over-explained or under-explained snippet of text.

DEDICATION

To the friends, families, and peers who supported me throughout the research process.

ACKNOWLEDGEMENTS

Contributors

I would like to thank my faculty advisor, Dr. Chaspari, and research associates, Ehsanul Haque Nirjhar, Md Nazmus Sakib, Raghu D. Veerappa, and Dr. Ani Nenkova for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

The data used for A Linguistic Analysis to Quantify Over-Explanation and Under-Explanation in Job Interviews were provided by the HUBBS lab at Texas A&M University. The analyses depicted in A Linguistic Analysis to Quantify Over-Explanation and Under-Explanation in Job Interviews were conducted in part by HUBBS lab at Texas A&M University and were published in 2022.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Undergraduate research was supported by the HUBBS lab at Texas A&M University. This study has been further supported by the U.S. National Science Foundation (NSF, Grant 1956021).

1. INTRODUCTION

1.1 Interpersonal Interactions

Interviews involving the potential for employment often include applicants' assessments through employers to fill a position at the company. When broken down to its most elemental components, these interviews encompass an interpersonal exchange amongst interviewers. To describe this process further, job applicants probe the interviewee for signals hinting at capabilities, behavior, skills, and likeability to determine a candidate best suited for the opening [2]. All factors and variables considered; this appears to be one of the most popular tools to achieve such a taxing responsibility. From the moment interviewees and recruiters meet for the first time, these "zero-acquaintance interactions" interviews are highly determinant on audio and visual behavioral cues related to both the applicant and the interviewer [4]. With this in mind, veterans are a group of Americans who have historically had less civilian work experience, taking a long time to find a job, particularly after leaving service [5]. Given, the constantly shifting environment encompassing the modern workplace, providing assistance to veterans has the potential to subdue the transition from military to civilian life.

1.1.1 Verbal and Non-Verbal Communication

When considering communication, two complementary methods to characterize interpersonal interactions are generally described as verbal and non-verbal. It's often noted that that non-verbal demeanor can be observed through the length of time spoken, timbre, and tone. Non-verbal behavior also inducts visual elements that often involve facial expressions, head motions, posture, and written responses [1]. From this, interviewers generally perceive and translate these social cues quickly with accuracy. This is likely designed from experiences, from

which is a result of an instinctive development that is typically challenging to simulate. On the other hand, verbal communication is accepted as a more primary mode of communication, where more obvious social variables including reputation, personality assessment, emotional capacity, and dominance are conventional outcomes with added nonverbal displays of behavior [28].

1.2 Approaches Using AI-Based Multimodal Analytics to Understand Interpersonal Interaction

More recently, the expanse of intelligent systems and cheaper computers have allowed for availability and improvements within the realm of visual and auditory sensors including video and microphones [21, 22]. This added to enhanced interpretative techniques allow for the categorically precise extraction of verbal and non-verbal behavioral patterns, are used in more modern surveys [24, 26]. With such data, machine learning techniques be leveraged in pursuit of effective algorithmic methods to generate automated inferences based on individual and group variables such as dominance traits, temperament, qualities of leadership, and attention span [23, 25, 27].

In the work presented, several solutions involving machine learning-based frameworks to are assessed to identify linguistic behaviors of interest in civilian job interviews. To set the stage, a mean interview duration lasting approximately 19.07 minutes, with a standard deviation of 6.78 minutes. From the collected data, video, audio, and self-reports from each participant were conducted and collected. Then, we proceeded to clean the data to suit a deep learning pipeline. Next, we used deep learning architectures and practices to automatically detect linguistic behaviors of interest during the interview. Compared to existing published papers, this work is the first to design machine learning analytics based on multimodal data to better understand strengths and challenges that U.S. veterans face during civilian job interviews. Despite not being

the focus of the paper, self-reports were recorded before and after the interviews, in addition to the visual footage. As a direct extension of the videos collected, a set of verbal interactions between both parties were transcribed into textual data sets.

Based on prior works involving multimodal data analytics of job interviews, lexical, prosodic, and linguistic aspects of the interviewee's replies where crystalized with the help of linguistic feature extraction [3]. The intended prediction model is designed to automatically obtains a distinct set of multimodal features (i.e. lexical, prosodic, and linguistic), and calculates the general interviewee's levels of explanation. Moreover, the relative feature correlations associated to verbal features ascertained by the applied regression models are investigated and used to measure relational significance in the context of explanation level. Notably, interviewee responses and associated elements in the pipeline are completely automated.

2. METHODS

2.1 Prediction Framework

For the prediction model, a selection of attributes from the verbal interviews were extracted automatically through a means of existing libraries and frameworks. Two deep learning algorithms were trained using the DistilBERT [6] a “distilled” or refined version of the RoBERT algorithm [7, 8]. The concept behind training these models are to reflect precise results and discover correlative pattens between definitive feature and established coherence scoring. This model framework is capable of supporting a variety of textual predictions methods which include text classification (binary and multiclass logistic regression), token classification, question answering, and text scoring among many others. Originally, the models are pretrained on a series of datasets [13, 14, 15, 16, 17]. However, as a proof of concept the Turker’s dataset [3] was used to train the predictive scoring, where the set of collected and scored interviewee interviews were later trained on a separate model for final assessment. Notably, these models are employed to measure and obtain an intuition on the comparative value of each attribute and the correlations discovered among them.

2.2 Feature Extraction

Of the collected data, two types of features for every interview were assessed; lexical and prosodic features to measure interviewees’ lexical and prosodic features with respect to the interviewers’ engagement. Both of the features selected are defined to reflect the behaviors that have been relevant in interview discussions (e.g., confidence, language substance, etc.). Supplementally, this is also founded on past literature on the basis of computationally recognized social behavior [3, 4]. It’s important to note that subtle social characteristics such as reputation,

personality assessment, emotional capacity, and dominance as aforementioned aren't as easily measured; verbal characteristics must be defined based on contingent attributes where social characteristics are then associated to an explanation label. When considering and evaluating consistent lexical characteristics, manually transcribed interviews accounted for all descriptive and nondescriptive features (which also perceived filler and meaningless words such as "hmm", "like", "uhh", and "umm") upon completion. From the results, a series of indexed lexical features and correlation scores were revealed from the set of interview snippets on a person-specific basis as seen in Figure 2.

2.3 Lexical Features

Based on the results of the final transcripts, the lexical features revealed valuable material regarding the prosodic correlation, personality quality, and interview substance. One of the greatest applied lexical attributes is the unigram quantities for each word. With this in mind, handling unigram totals as features frequently produces a thinly populated, yet high-dimensional feature vectors, even with a constrained amount. As a result, this challenge is addressed using a couple of methods. To start, rather than of using crude unigram counts, we employed counts of numerous word categories defined by psycholinguistic syntax, called "Linguistic Inquiry Word Count" (LIWC) [10]. The LIWC software relies on a pre-defined dictionary in which words correspond to grammatic, syntactic, socio-emotional, and cognitive categories. Examples of these categories include positive (such as kind, happy), functionally distinct word groups (involving conjugates, articles), and negativity (including angry, sad). Feature selection is explored with an expensive backward elimination approach, starting with the LIWC features, and eliminates a single feature at a time so its deduction results an optimal gain in accuracy while also accounting for cross-validation. This method is continued until any additional feature

elimination negatively or ceases to effect accuracy. As a result, approximately 23 features using LIWC were selected. The correlation between categories with various psychological traits generally offer significant traits about skills associated with social and personality. Hence, several of these groupings are associated to the overarching performance of the interview. To tie the datasets together, each of the LIWC features were amalgamated to the existing data frame where a set of trait scores are correlated to the presented snippet of text.

2.4 Semantic Analysis

Despite the fact that the LIWC feature extrapolation is referenced as a sturdy medium for describing text, it's important to note that it has a set of constraints that limit the functionality of the corresponding scoring labels. More specifically, LIWC uses a dictionary which distinguishes words, and assess whether they are appropriately fitted to a predestined classification. For example, the words good, appreciate, and pleasing are mapped to the class titled "positive emotion", and by means of cataloging these phrases, LIWC intends to deliver a suggestion pertaining to how positive the text conveys. In typical communication, good does not automatically reveal a particularly positive reaction, but could also be impartial or possess a negative connotation when the subject expresses irony. To combat this issue, we used the DistilBERT model to quantify semantic dimensions of interviewees' responses.

When considering the DistilBERT model, the configuration parameter, "max_position_embeddings" possesses a limitation where up to 512 embedded tokens can be processed, signifying inputs are limited by length. Moreover, token indices with a dimension extending past the constrained maximum sequence length for this model must be further broken down or truncated. Hence, running a model containing a sequence larger than the specified limit will cause in indexing faults [8]. To address this, truncation was enabled with padding the

maximum length of acceptable input length (512). Ultimately, this resulted in truncating each of the tokens, where a given item from the longest set in the pair batch of pairs or sequences is eliminated. As a result, a heuristic truncation can be achieved.

3. RESULTS

3.1 Results & Discoveries

3.1.1 Data Collection & Processing

Data collection methods are performed in a recorded virtual setting between an interviewee and a company representative interviewing applicants for a tangible role. It's important to note that this was a mock interview was not intended for actual hiring. Participants are asked to wear a heartbeat monitor placed over the chest, Fitbit wristband, and were visually and auditorily recorded for the duration of the interview. Upon completion of the interview, interview audio files were manually translated to text with their respective timestamps using Sneedacity [9]. Once complete, feature extraction was performed using LIWC text analysis program to calculates the extent to which different groups of words are used in a transcript. This metric is considered the de facto standard and is capable of measuring texts involving copied natural language documents in various text formats.

3.1.2 Data Exploration & Feature Extraction

Prior to performing any formal methods of natural language processing, it's important to get an understanding of what the data looks like. The primary motivation for this is to be conscious of potentially inherent biasing or irregular distributions hidden in the data. Of the collected interviews, the interviewees were significantly biased towards males as seen in Figure 1. Two significant factors to account for when exploring the datasets are the duration of speech and rating of spoken patterns. Abstractly speaking, disproportionate text length introduces a new set of hyper-parameters or needed workarounds in the applied Natural Language Processing (NLP) model [12]. These features or workarounds are not essential to NLP model but exist to

only handle uneven text lengths. Based on the set of participants, samples of text from their responses were used to assess the proper explanation label associated with the quality of their responses. Additionally, using the LIWC processed transcripts feature correlations associated with considered outcomes were examined for further model assessment. In particular, Figure 3 describes the general heatmap describing the associative trends between LIWC features. These traits generally described correlative capacities including analytic, confidence, leadership, prose, performance, perception, time consciousness, and individual and team-based pronoun association as described in Figures 4 to 11 respectively. Each of which are useful in finding correlated and uncorrelated behavior, which could effectively map to a higher or lower explanation label. In addition to the feature extraction performed over the text, verbal coherence was measured in a similar manner on a scale from 0 to 3. A range of favorable and unfavorable responses were identified where 0 is represented as an underexplained description, 1 signifies a succinct response, 2 illustrates a comprehensive reaction, and 3 is defined as an overexplained retort. Particularly, the ratings, 0 and 3 are identified as unfavorable on extreme ends of the spectrum, where the interviewee's responses failed to hit all the key points of a question asked from the interviewer. Moreover, 0 is a response that fails to precisely address the question and is often brief, whereas 3 delivers a proper response, but will lead on to repeat previously mentioned statements to a fault. More favorably, 1 and 2 are more amiable responses that successfully address the question directly addressed to the interviewee. To expand, succinct responses labeled as 1 possess key points that are briefly and clearly expressed. Similarly, comprehensive responses are exemplified by a reaction to question that is completely answered, involving all or virtually all elements or aspects of the subject. Based on the collection of data, a total of 165 responses were collected from the collection of interviews. In particular, 16 underexplained, 72

succinct, 60 comprehensive, and 17 over explained replies as illustrated in Figure 2. Clearly, a heavy bias leaning towards more favorable responses were collected as opposed to less favorable reactions.

3.1.3 Model Training & Design

This design of the model is heavily influenced by a "distilled" edition of the RoBERT-base NLP model. It obeys similar training procedures as previously used in the DistilBERT design. Notably, the model has a total of 12 heads, 768 dimension and, 6 layers resulting in 82 million configurable parameters.

By design, RoBERT is a model based on the transformer structure from which it is typically a pretrained self-supervised model based on a substantial number of records in English. However, for the purpose of the model builds upon the pretrained English records by retraining the model to recognize multi-class sets associated to level of explanation. As a result, this implies it was initially pretrained on natural text, without manual categorization, but also appends the newly trained label associations to interviewee level of explanation. Consequentially, the model possesses an automated method to produce responses and their respective labels from scripts.

Moreover, the base model was already trained with the goal of Masked Language Modeling (MLM) scoring [11]. As a result of accepting a sentence, the NLP model arbitrarily screens 15 percent of the input texts and proceeds to manage the complete masked set of words over the model and finally predicts the screened phrases. This is particularly unique from conventional Recurrent Neural Networks (RNNs) such that the words are seen from models that champion autoregression such as the Generative Pre-trained Transformer (GPT). This updated technique ultimately lets the NLP model learn a sentence representation bidirectionally.

Hence, the NLP model can understand an internal description of a language which can be applied to root out features suitable for tasks encountered on a test set. In simple terms, this approach leverages the labeled sentences to train a classifier using the features created by the NLP BERT architecture as feedback. To expand, the model is mainly intended as a fine-tune on categorizers that use an entire sentence to produce a result (structure classification). The model is trained using cross-fold validation with interviewee partitions. The primary goal of this approach is to not only achieve and reasonably optimal test accuracy and loss, but to also recognize the effects of common speech patterns.

To further analyze the results, a similar approach to achieving a more favorable accuracy and loss came in the form of a binary logistical regression, which is applied to the model. By specifically targeting adjacent text labels (i.e. 0 and 1, or 2 and 3), a finer scale of accuracy can be assessed to understand the classification of verbal speech patterns based on definitive features between the compared groups. It's important to note that this particular method is also built on top of the DistilBERT model, so a similar training, validation, and testing can take place. The model structure is sequentially designed beginning with a linear neural network of dimensions (768, 768), ReLu, dropout, and ending with a linear neural network of dimensions (768, 4). Here, a tokenizer is employed with a maximum sequence length of 512 where the initial 510 tokens are used for classification. Finally, the fine-tuning parameters are designed such that the batch size is 16, with 20 epochs and learning rate of 1e-05. Much like the previous technique, this method is trained employing the same methods involving randomized cross-fold validation with interviewee partitions. With this new approach, the similar, yet comparable responses produced significantly higher results compared to the original method applied. In essence, underexplained

text can be differentiated from succinct replies while, comprehensive answers are categorized apart from over explained content.

3.1.4 Model Results

Starting with the initial model, the multiclassification DistilBERT model resulted in with an accuracy and RMSE loss of 68.2 percent and 3.02 respectively over the test set. Moreover, the precision range approximately from 0.126 to 0.306. The training process is illustrated in Figure 12, where the applied predictions are depicted in Figure 13. Despite the fact that this is an improvement from a randomized partitions irrespective of the interviewee, this achieves reduced performance in results which are likely due uncorrelatable features between the wholistic set of interviewee responses and the high biasing towards an explanation label of 1 (or succinct). To expand, further, we can observe in Figure 14, a sporadically colored confusion matrix, which confirms the model's low performance.

On the other hand, the DistilBERT logistical binary classification technique using cross-fold validation between explanation labels 0 and 1 (underexplained, and succinct) rating resulted in with a test accuracy of roughly 82.9 percent, the loss reaches about 0.31, and possesses a precision range approximately from 0.333 to 0.818. The training and prediction results can be described in Figures 15 and 16 respectively. On the other hand, the same method applied to the comprehensive and overexplained ratings resulted in a test accuracy of around 81.3 percent, a loss of about 0.333, and maintains a precision range from 0.667 and 0.833. The training and prediction findings can be expressed in Figures 15 and 18 respectively. Based on binary logistic regression, Figures 17 and 19 highlight densely and orderly populated confusion matrices. Hence, describing more favorable results between the enhanced DistilBERT models.

Based on the findings of the multi-class classification and binary classification DistilBERT models, we can clearly observe a significant improvement in the quality of results when comparing each simulation. As a result, the best performing model is the DistilBERT binary classifier model which effectively captures a higher accuracy and lower losses between both label groups.

4. CONCLUSION

4.1 Concluding Remarks

In the thesis, a computational basis for characterizing important interview behaviors to the effect of job interviews is proposed using interviewees' verbal and nonverbal indicators assessed from the auspices of visual and audio cues.

From the recorded interactions, verbal attributes were then obtained from the interviewer and applicant interactions and behavior. Internally, correlation analysis was performed between the difference linguistic measures. Following this step, semantic analysis enabled by the DistilBERT model attempted to estimate the level of explanation of an interviewees' answer.

4.2 Future Work

It's important to account for supplementary factors that must be considered with interviewees. Particularly, exploring a set of additional social and interpersonal factors would likely establish a feedback system that can also account for these considerations. By introducing NLP model biasing to account for relevant internal factors such as somatization, interpersonal sensitivity, depression, and obsessive-compulsive disorders have the possibility to point to more accurate results. While performing data exploration, these set of factors were heavily considered and were analyzed based on a subsample of participant in order to poll for an identifiable correlation. As an alternative set of avenues to continue the findings, incorporating vocal features to the linguistic measures and examining the interplay of conversation between the interviewer and the interviewee could reveal a new set of dynamics for further analysis. These approaches can suggest that further NLP model design and feature extraction could be leveraged to properly

account for the described attributes and limiting factors. Hence, these advances can be assessed for additional research.

REFERENCES

[1] Hemamou, L., Felhi, G., Martin, J.-C., & Clavel, C. (2019). Slices of attention in asynchronous video job interviews. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). https://doi.org/10.1109/acii.2019.8925439

[2] Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J.-M., Schmid Mast, M., & Gatica-Perez, D. (2016). Training on the job: Behavioral analysis of job interviews in hospitality. Proceedings of the 18th ACM International Conference on Multimodal Interaction. https://doi.org/10.1145/2993148.2993191

[3] Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. IEEE Transactions on Affective Computing, 9(2), 191–204. https://doi.org/10.1109/taffc.2016.2614299

[4] Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Transactions on Multimedia, 16(4), 1018–1031. https://doi.org/10.1109/tmm.2014.2307169

[5] Gonzalez, J. A., & Simpson, J. (2021). The workplace integration of veterans: Applying diversity and fit perspectives. Human Resource Management Review, 31(2), 100775. https://doi.org/10.1016/j.hrmr.2020.100775

[6] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March 1). Distilbert, a distilled version of Bert: Smaller, faster, cheaper and lighter. arXiv.org. Retrieved April 2, 2022, from https://arxiv.org/abs/1910.01108v4

[7] Roberta: A robustly optimized Bert pretraining approach. (n.d.). Retrieved April 2, 2022, from https://www.cs.princeton.edu/~danqic/papers/roberta_paper.pdf

[8] Roberta-base · hugging face. roberta-base · Hugging Face. (n.d.). Retrieved March 7, 2022, from https://huggingface.co/roberta-base

[9] Sneedacity. (n.d.). Retrieved March 7, 2022, from https://sneedacity.org/

[10] LIWC. Welcome to LIWC-22. (n.d.). Retrieved March 7, 2022, from https://www.liwc.app/

[11] Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2021, January 1). Masked language model scoring. arXiv.org. Retrieved March 7, 2022, from https://arxiv.org/abs/1910.14659

[12] STIGEBORN, OLIVIA (n.d.). Text ranking based on semantic meaning of sentences. Retrieved April 2, 2022, from https://kth.diva-portal.org/smash/get/diva2:1589461/FULLTEXT01.pdf

[13] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 2015 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2015.11

[14] Wikimedia Foundation. (2022, March 26). *English Wikipedia*. Wikipedia. Retrieved April 2, 2022, from https://en.wikipedia.org/wiki/English_Wikipedia

[15] Nagel, S. (2016, October 4). Common crawl. Retrieved April 2, 2022, from https://commoncrawl.org/2016/10/news-dataset-available/

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language models are unsupervised multitask learners*. Retrieved April 2, 2022, from https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[17] Trinh, T. H., & Le, Q. V. (n.d.). *A simple method for commonsense reasoning*. Retrieved April 2, 2022, from https://teaching.bb-ai.net/Student-Projects/Winograd-Challenge-Papers/Trinh2018-Simple.pdf

[18] Balaji, S., Gopannagari, M., Sharma, S., & Rajgopal, P. (2021). Developing a machine learning algorithm to assess attention levels in ADHD students in a virtual learning setting using audio and video processing. *International Journal of Recent Technology and Engineering (IJRTE)*, *10*(1), 285–295. https://doi.org/10.35940/ijrte.a5965.0510121

[19] Rello, L., & Ballesteros, M. (2015). Detecting readers with dyslexia using machine learning with eye tracking measures. *Proceedings of the 12th International Web for All Conference*. https://doi.org/10.1145/2745555.2746644

[20] Wang, L.-J., Lee, S.-Y., Tsai, C.-S., Lee, M.-J., Chou, M.-C., Kuo, H.-C., & Chou, W.-J. (2019). Validity of visual and auditory attention tests for detecting ADHD. *Journal of Attention Disorders*, *25*(8), 1160–1169. https://doi.org/10.1177/1087054719887433

[21] Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., & Sabouret, N. (2013). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8253 LNCS*, 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

[22] Nguyen, L. S., & Gatica-Perez, D. (2016). Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, *18*(7), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

[23] Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. (2018). Automated video interview judgment on a large-sized corpus collected online. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, *2018-Janua*, 504–509. https://doi.org/10.1109/ACII.2017.8273646

[24] Hoque, M. E., Courgeon, M., Martin, J., Mutlu, B., & Picard, R. W. (2016). MACH: My Automated Conversation coacH Mohammed. *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[25] Hartholt, A., Mozgai, S., & Rizzo, A. (2019). Virtual job interviewing practice for high-anxiety populations. *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 238–240. https://doi.org/10.1145/ivade780

[26] Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.-C., & Clavel, C. (2019). HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 573–581. https://doi.org/10.1609/aaai.v33i01.3301573

[27] Batrinca, L., Stratou, G., Shapiro, A., Morency, L. P., & Scherer, S. (2013). Cicero - Towards a multimodal virtual audience platform for public speaking training. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8108 LNAI*, 116–128. https://doi.org/10.1007/978-3-642-40415-3_10

[28] Chollet, M., Wörtwein, T., Morency, L. P., & Scherer, S. (2016). A multimodal corpus for the assessment of public speaking ability and anxiety. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 488–495.

APPENDIX: A

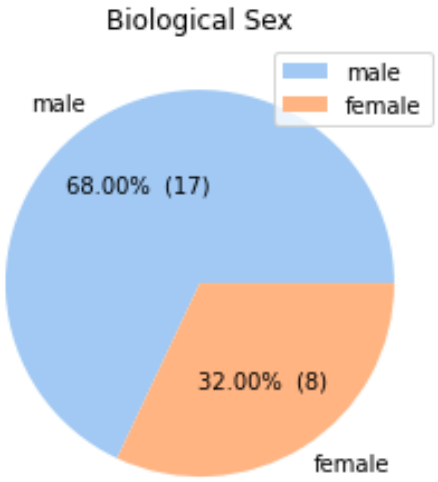

Figure 1; Participant Demographic by Sex

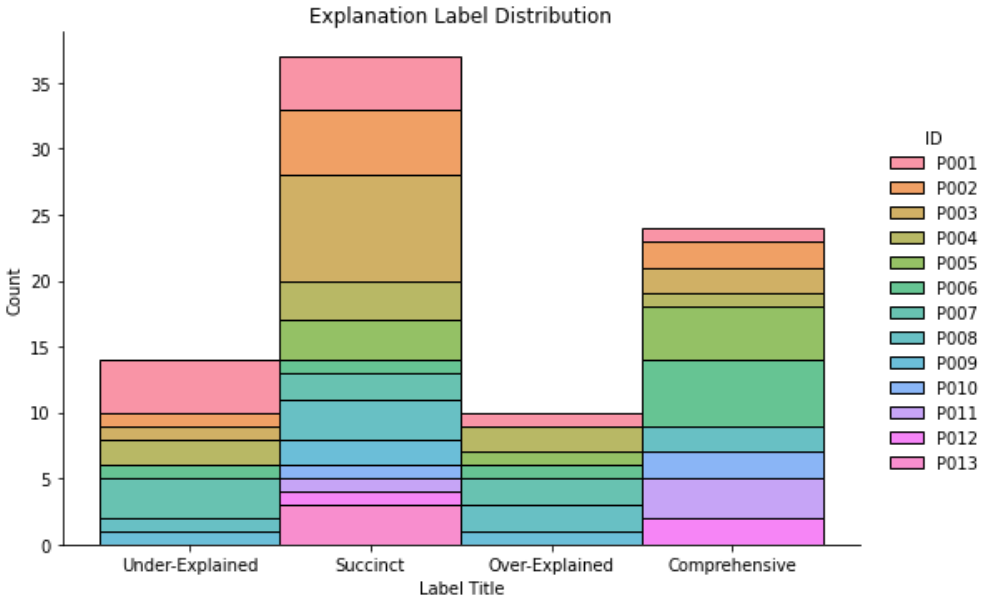

Figure 2; Explanation Label Distribution by Participant

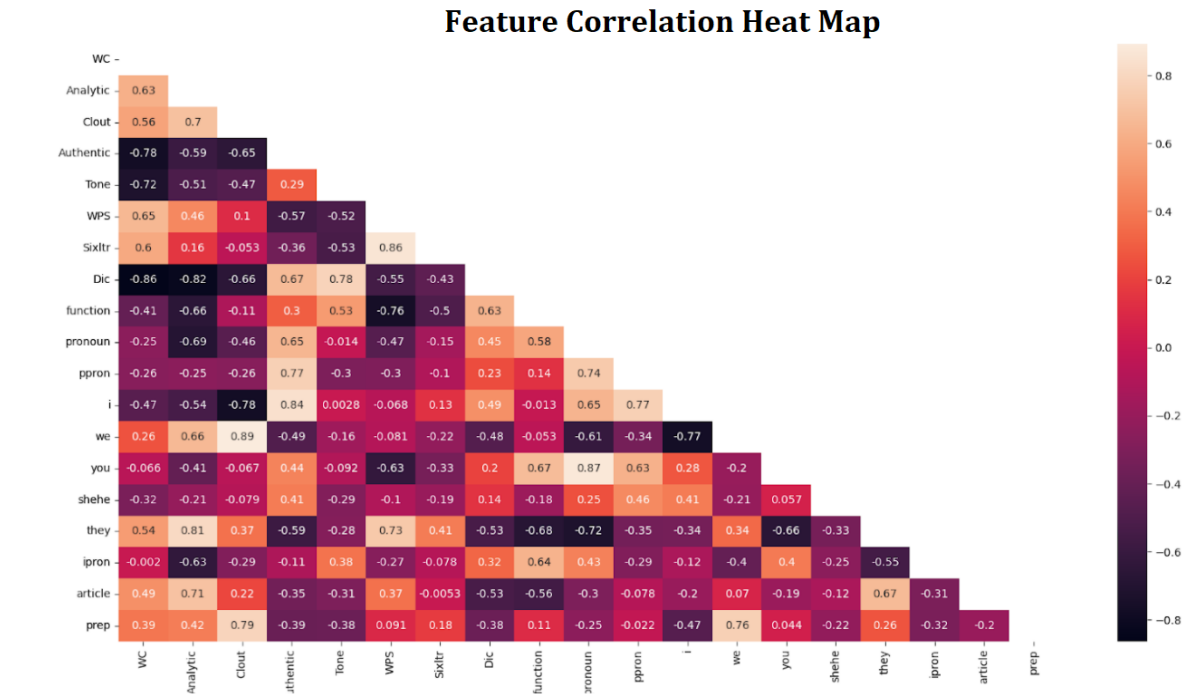

Figure 3; Participants' LIWC Feature Correlation Heat Map

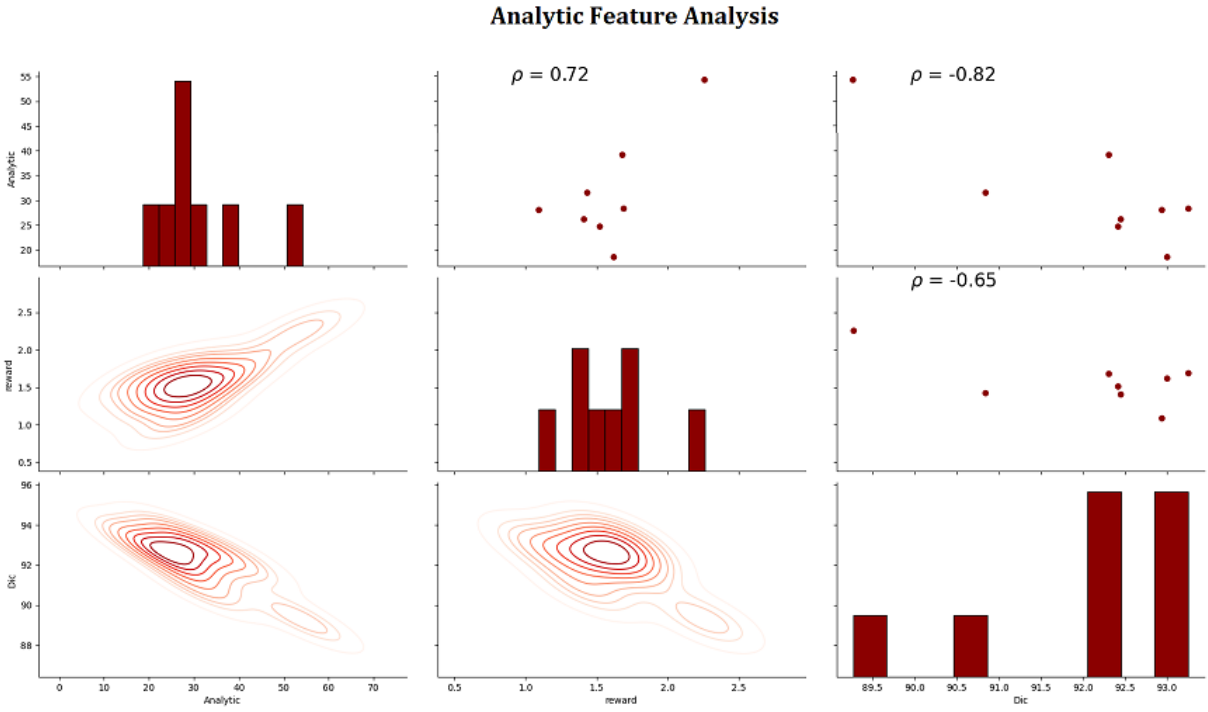

Figure 4; Dictation, Reward, Analytic Feature Analysis

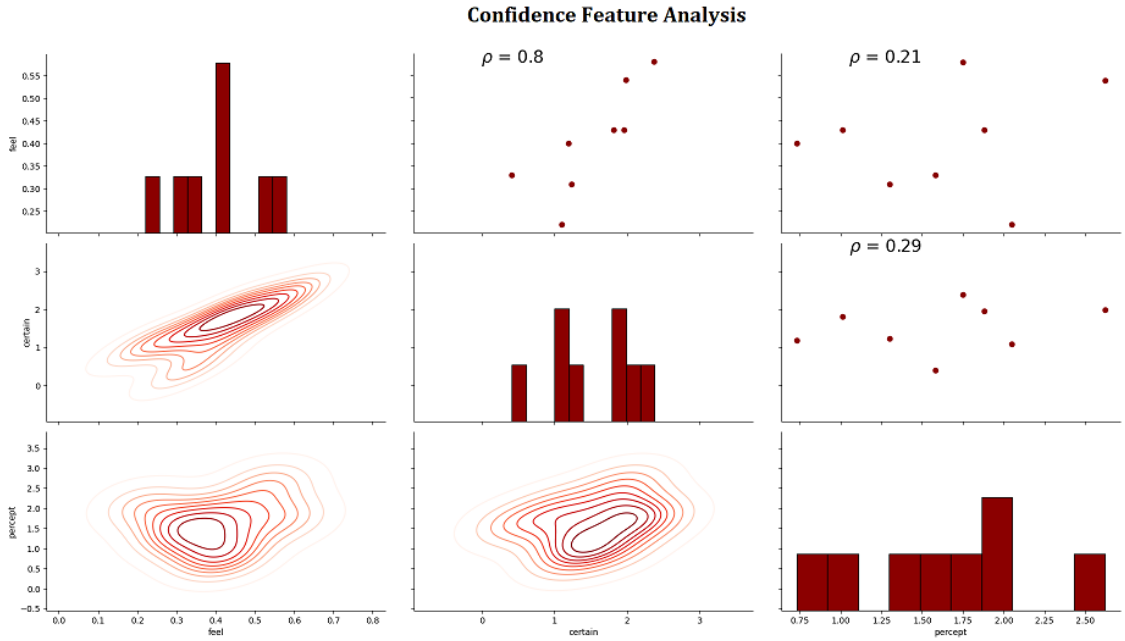

Figure 5; Perception, Certainty, Feeling Feature Analysis

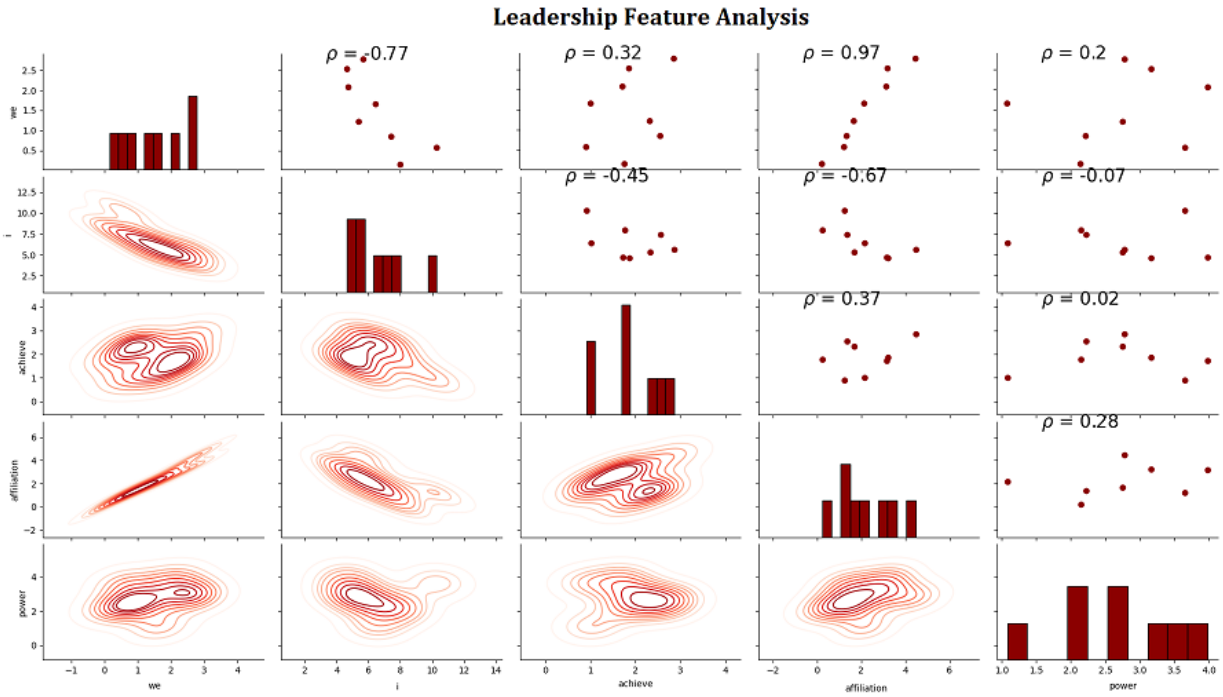

Figure 6; We, I, Achievement, Affiliation, Power Feature Analysis

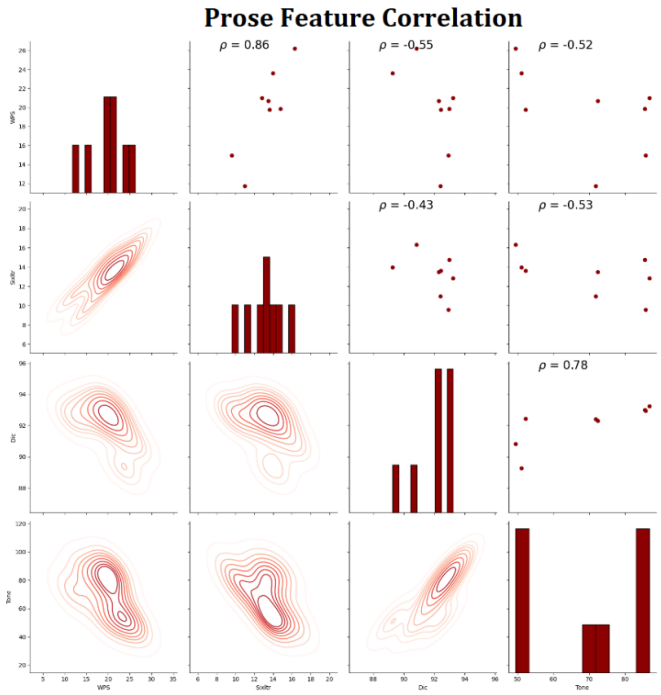

Figure 7; Tone, Dictation, Words Containing Over 6 Letters, Words Per Sentence Feature Analysis

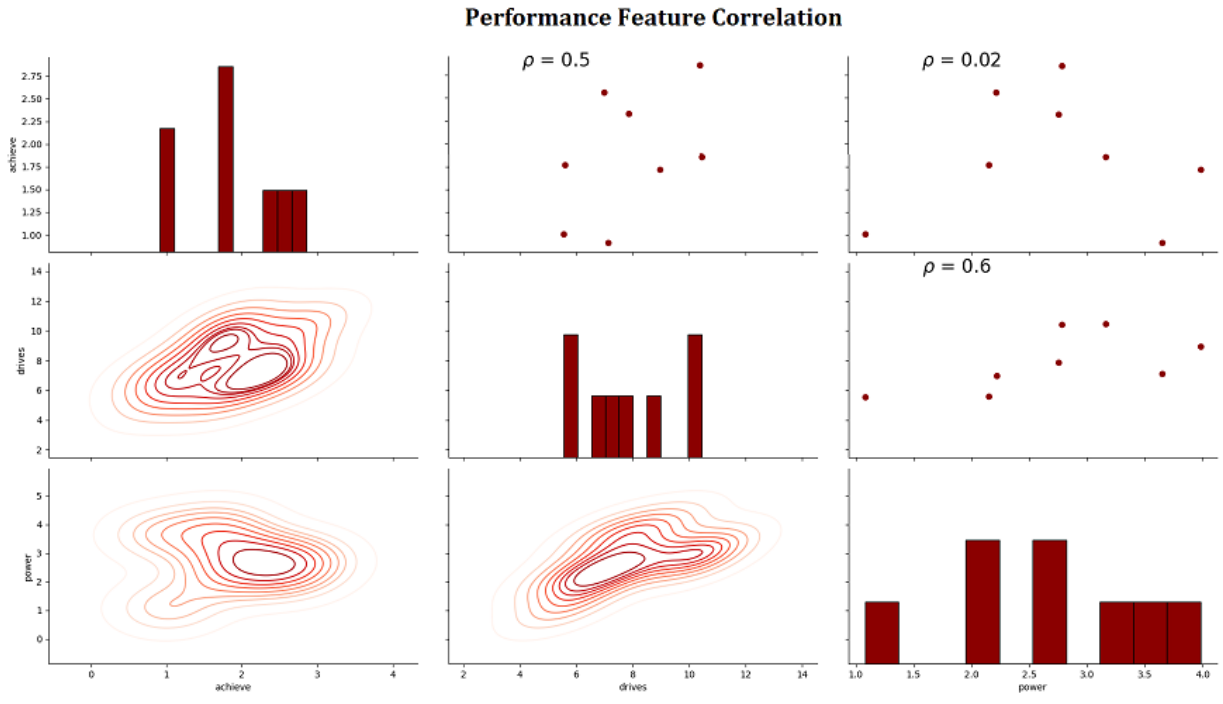

Figure 8; Power, Drives, Achievement Feature Analysis

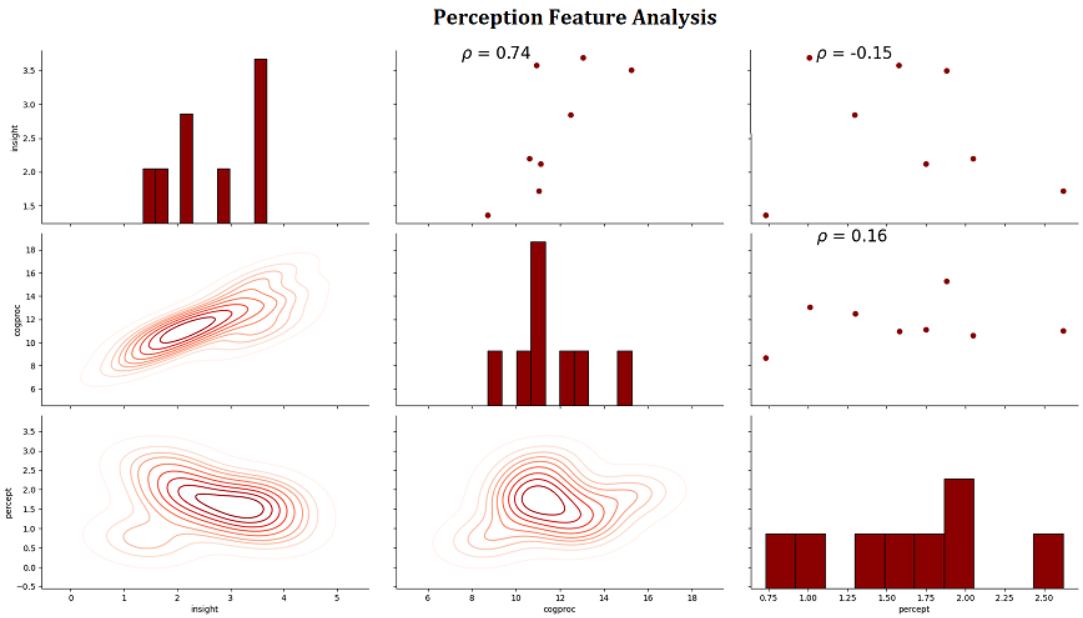

Figure 9; Perception, Cognitive Processes, Insight Feature Analysis

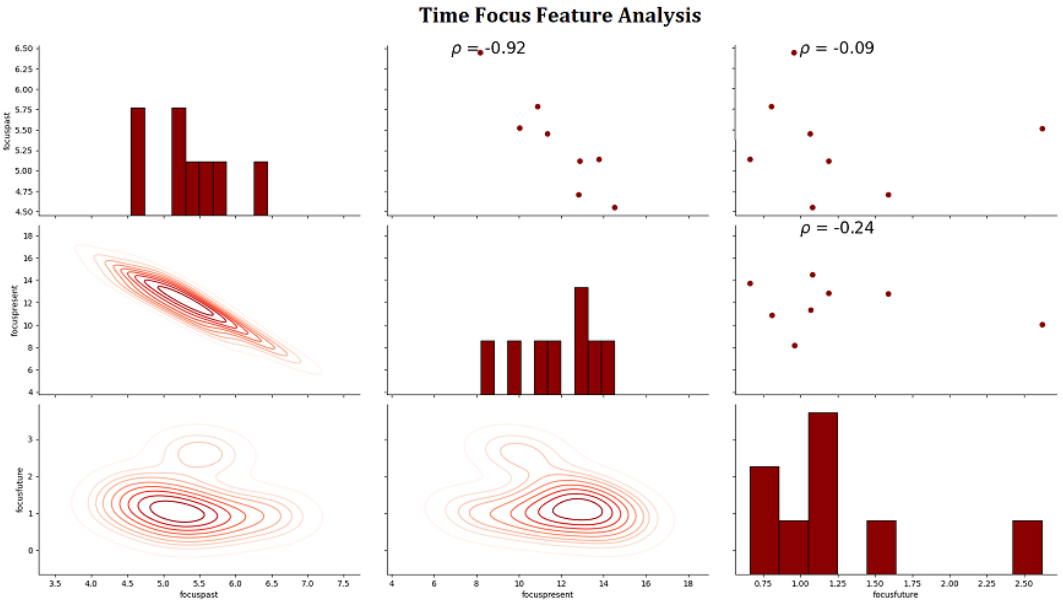

Figure 10; Future, Present, Past Focus Feature Analysis

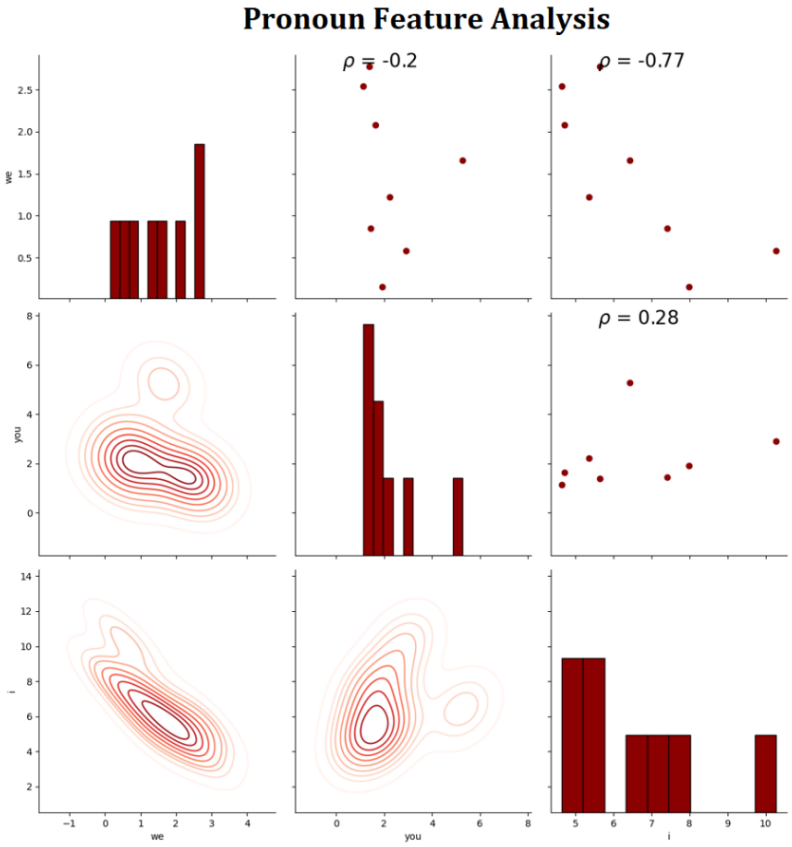

Figure 11; I, You, We Feature Analysis

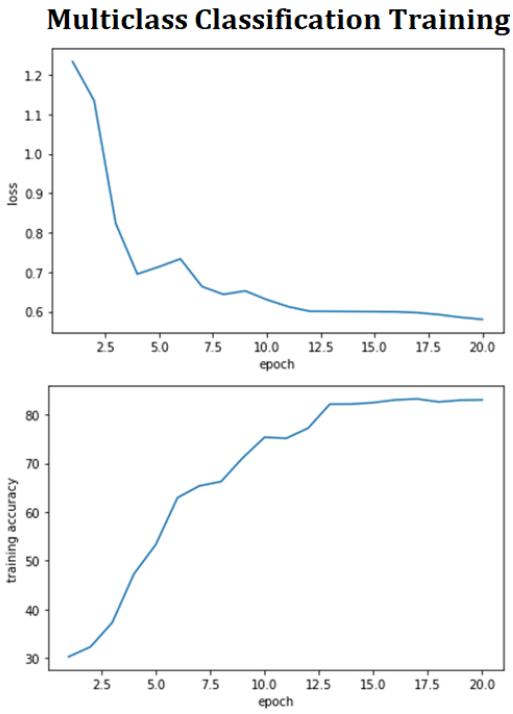

Figure 12; Multiclass Classification Training Accuracy and Loss

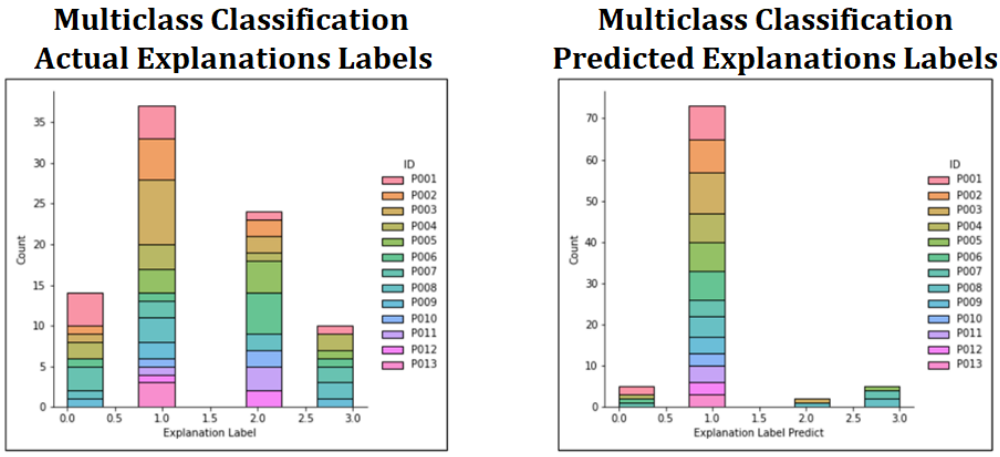

Figure 13; Actual vs. Predicted Multiclass Classification Explanation Labels

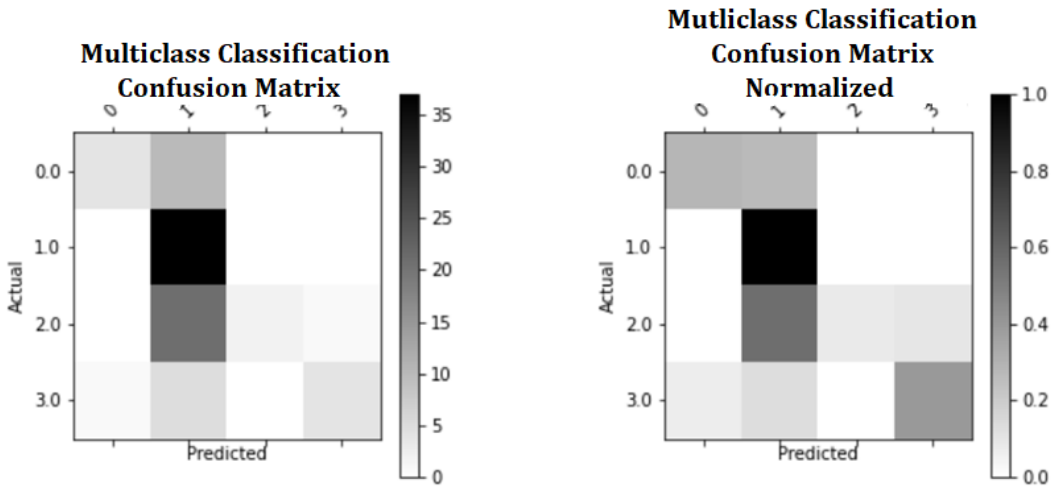

Figure 14; Non-Normalized vs. Normalized Multiclass Classification Matrices

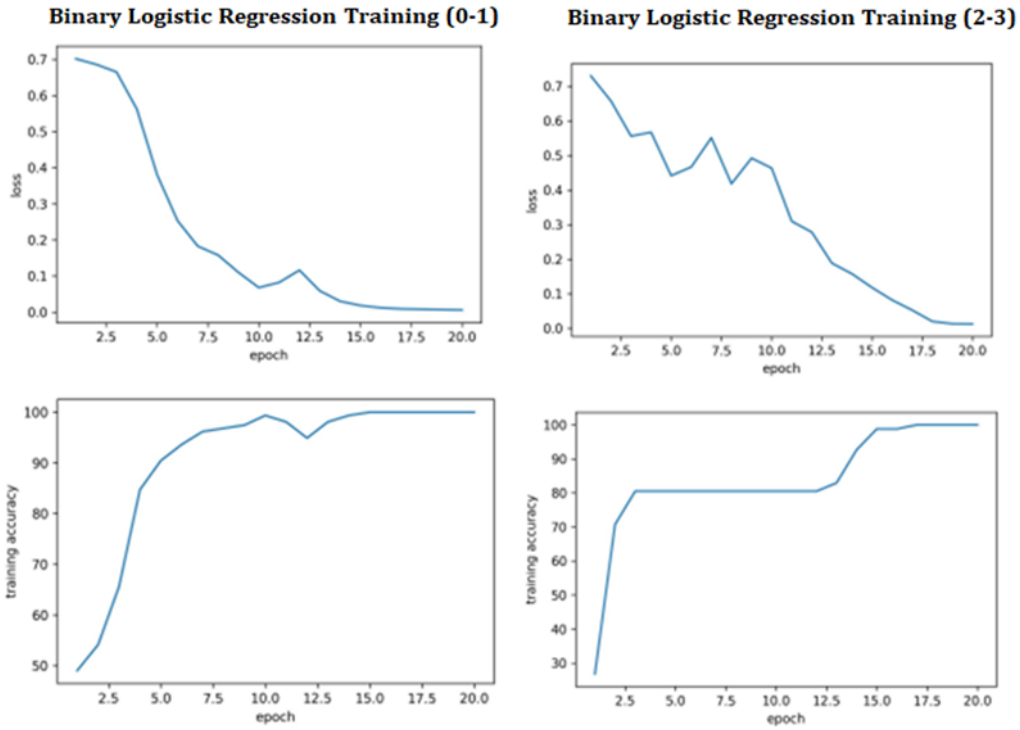

Figure 15; Binary Logistic Regression Training Accuracy and Loss

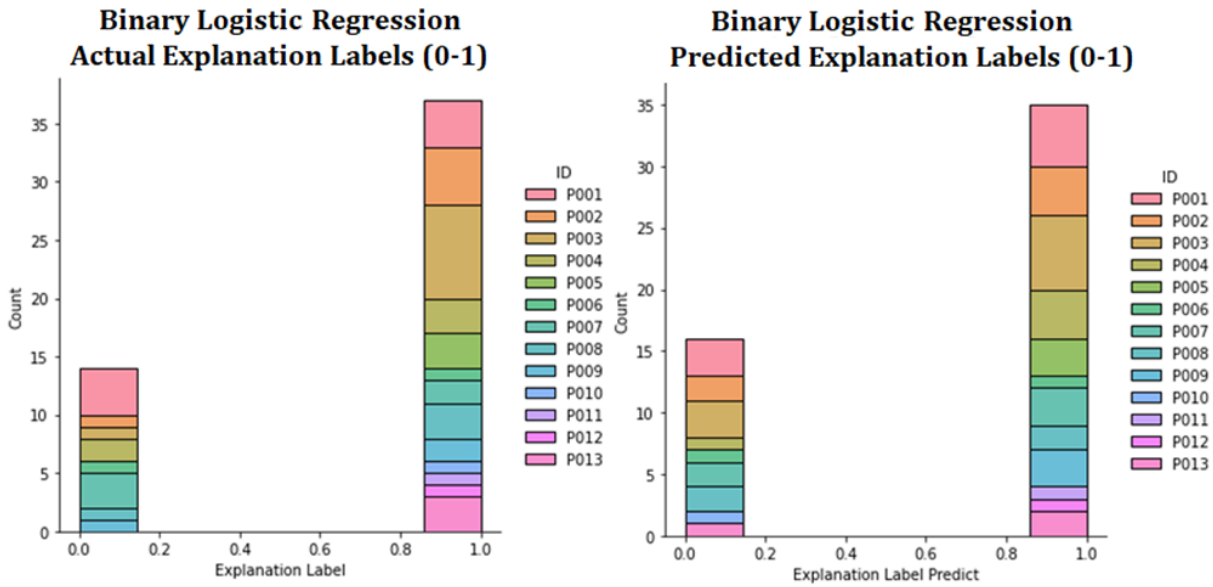

Figure 16; Actual vs. Predicted Binary Logistic Regression Explanation Labels (Under-Explained vs. Succinct)

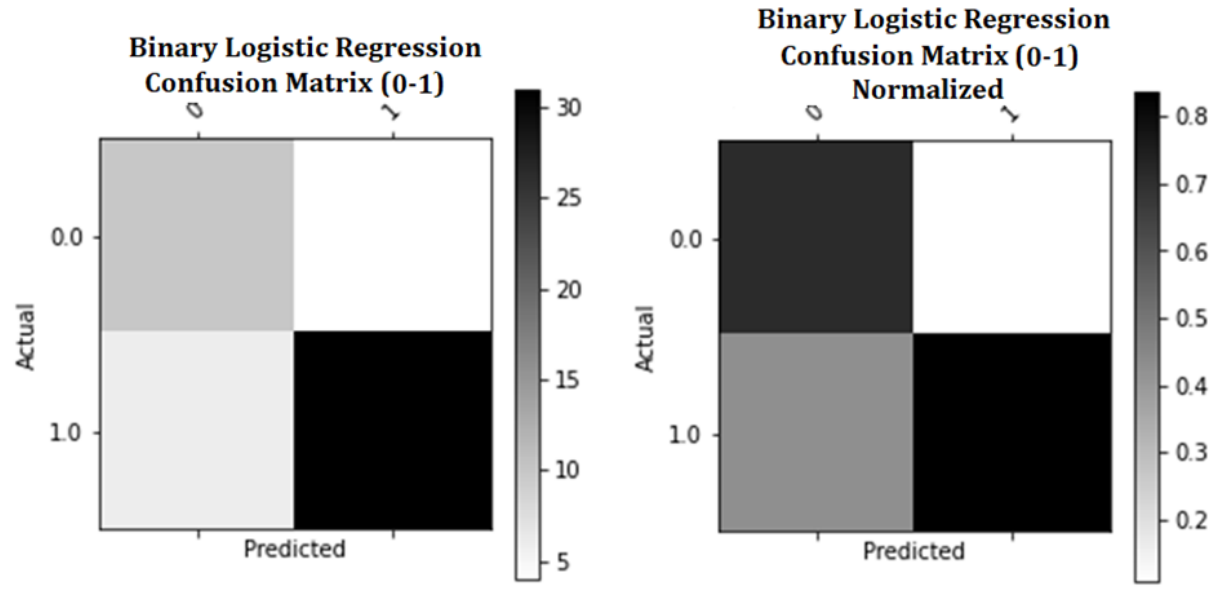

Figure 17; Non-Normalized vs. Normalized Binary Logistic Regression Matrices (Under-Explained vs. Succinct)

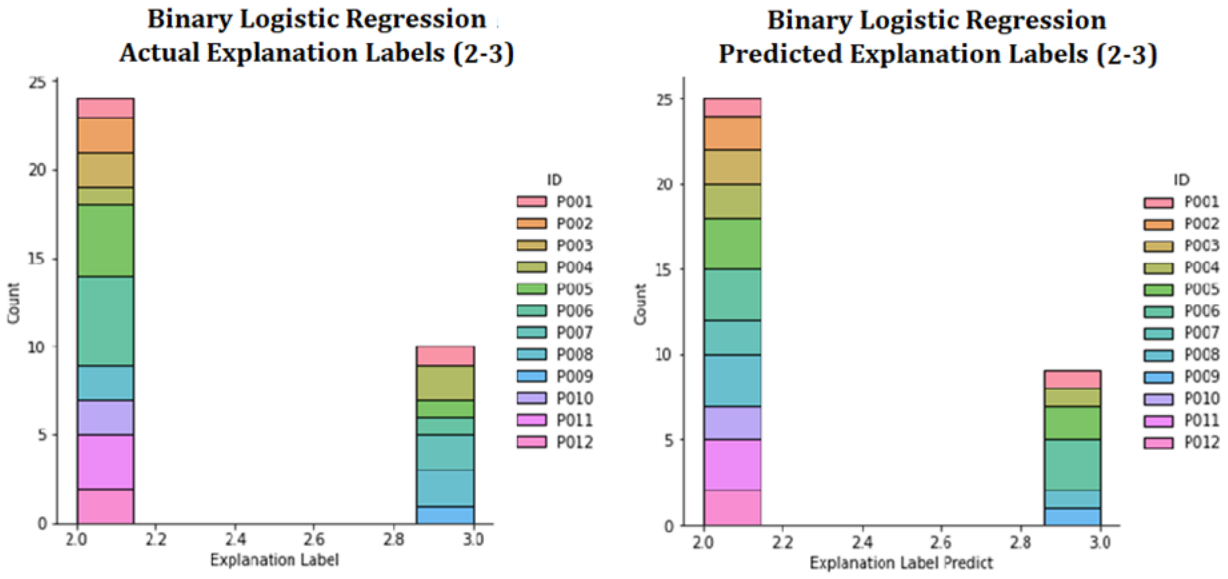

Figure 18; Actual vs. Predicted Binary Logistic Regression Explanation Labels (Comprehensive vs. Over-Explained)

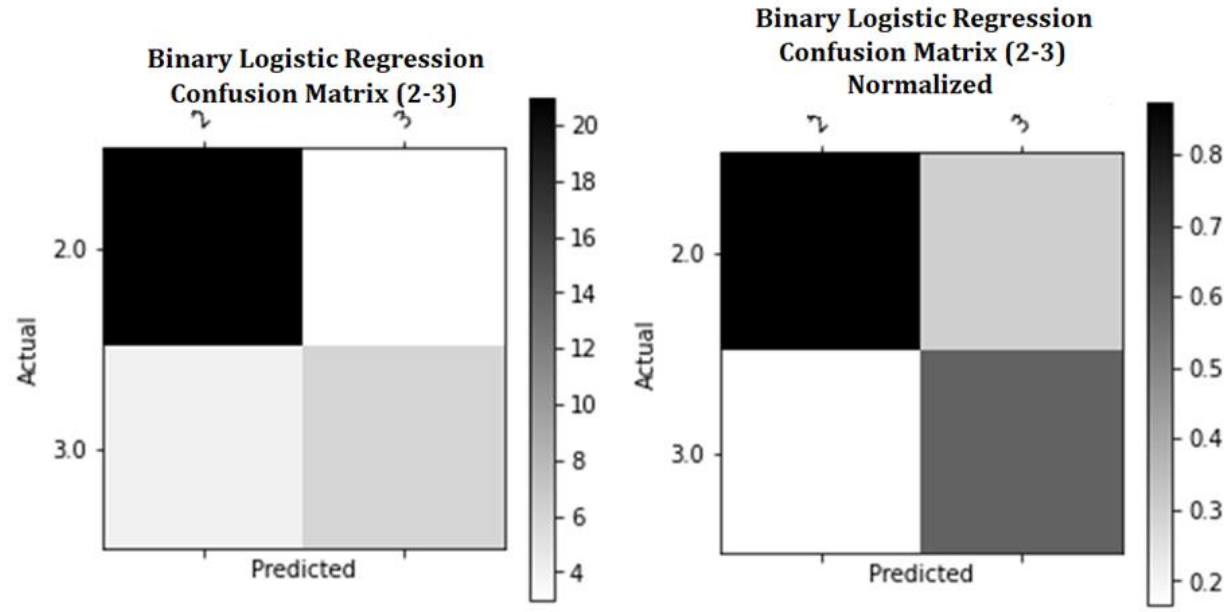

Figure 19; Non-Normalized vs. Normalized Binary Logistic Regression Matrices (Comprehensive vs. Over-Explained)