CSCI 5622
Homework 4: Designing ML models for real-world problems

Instructions for homework timeline and submission

1. In this homework you will work within teams of 4 classmates. The team assignments are uploaded on CANVAS. You can reach out to your team members via CANVAS.

2. Each team is randomly assigned to one of the three studies. Please see assignments on CANVAS.

3. You will present your work with your team in class on **December 2, 4, and 6** during class time (**10.10-11pm MT**). The assigned time slot for your team has been announced on CANVAS. Most of the work should be ready by the class presentations, including the data analysis and the presentation. **Participation in the presentation for all three days (December 2, 4, 6) is mandatory will count toward the in-class participation grade. Everyone will be called on by name to ask questions during the presentations. Each team will have 4 minutes to present, followed by 4 minutes of questions.**

4. The final report is due on **December 13, 2024 @ 11.59pm**. Please create a zip file with two pdf files, including the final report and the presentation. **One member per team can make the submission.**

5. You can use any publicly available library or code for this homework.

6. The total for this homework is **14 points** (out of 100 total for the class) and **2 Bonus points**.

Scope of Work

Language-based machine learning (ML) technologies are transforming digital health and education by enabling systems to process, understand, and generate human language effectively. In digital mental health, these technologies assist in diagnosing, monitoring patient treatment by analyzing conversational data. In education, these systems can serve as a foundation to intelligent tutoring systems that can track the learners' responses, assess performance, and provide personalized feedback.

This homework demonstrates the use of these systems in mental health and job interview training via three studies. It further examines ways to make these technologies more responsible via increasing their explainability, mitigating potential socio-demographic bias, and enhancing data privacy.

Study 2: Designing explainable speech-based machine learning for the estimation of job interview outcomes

Automated interview evaluation systems can play a valuable role in interview training by simulating real interview conditions and providing quantitative feedback to candidates on various aspects of their performance. These systems can assess verbal and non-verbal cues, such as tone of voice, pacing, facial expressions, and word choice, to identify strengths and areas for improvement.

This study examines the ability of speech-based ML models, that rely on both language and prosody, to automatically estimate interview outcomes, such as the interviewee's overall performance and excitement. The data come from the MIT Interview dataset [3, 5]. The file 'transcripts.csv' includes the interview transcripts. Each participant has conducted two interviews. The first column corresponds to the participant ID, x and whether the transcript belongs to the first interview (i.e., 'px') or to the second interview (i.e., 'pp$x$'). The prosodic features for each participant are found in file 'prosodic_features.csv'. The file 'scores.csv' contains scores for each interview, covering both interviewee's overall performance and level of excitement, as perceived by a third-party annotator. These scores range from 1 to 7 and will serve as the outcomes of the ML models.

Randomly split the participants into 5 folds and report results accordingly in the following experiments.

(a) (2 points) Extracting language features. Extract several language features from the data. Include at least two of the following types of features, spanning different levels of complexity. At least one of the two types of extracted features should be interpretable by humans.

- Syntactic vectorizers: count vectorizer (e.g., *CountVectorizer* from sklearn) transforming a collection of text documents into a numerical matrix of word or token counts; TF-IDF vectorizer (e.g., *TfidfVectorizer* from sklearn) incorporating document-level weighting, which emphasizes words significant to specific documents' part-of-speech features counting the distribution of part of speech tags over a document
- Semantic features: sentiment scores (e.g., Vader, `https://github.com/cjhutto/vaderSentiment`), topic distribution (using topic modeling), or named entities
- Advanced features: word embeddings, such as Word2Vec or BERT (e.g., *pytorch-pretrained-bert*)) for capturing contextual meaning

Below are some additional resources that could be useful for feature extraction: NLTK toolkit (`https://www.nltk.org/`); Google word2vec (`https://code.google.com/archive/p/word2vec/`); Hugging Face (`https://huggingface.co/`); Jurafski & Martin, Speech & Language Processing, Chapter 6: Vector Semantics and Embeddings (`https://web.stanford.edu/~jurafsky/slp3/6.pdf`).

(b) (2 points) Language feature selection. Explore a filter feature selection method of your choice to identify the k features belonging to the interpretable feature set from question **(a)** that are the most relevant to the two considered outcomes. Please discuss your findings and how these measures can be used to provide actionable insights to the user.

Note: Along with discussing the strength of the association between each feature and each outcome, please also comment on the direction of this association (i.e., whether it is positive or negative).

(c) (2 points) Estimating interview outcomes based on language. Use one tree-based ML and one deep learning ML algorithm of your choice to estimate the level of interview performance and excitement that was rated for each participant. You can use your findings from question **(b)** to determine the feature set. Please report the Pearson's correlation r and absolute relative error (RE) between the estimated and actual scores. Experiment with different values of k from question **(b)**. Please discuss your findings (e.g., Is this performance acceptable for real-world applications? What is the computational cost of the ML models and can they be deployed for edge applications?).

Note: The absolute relative error, RE, is defined as follows:
$RE = \frac{|\text{estimated PHQ-8} - \text{actual PHQ-8}|}{\max(\text{PHQ-8})}$

(d) (2 points) Multimodal ML models. While the content of what people say is important, how they say it is equally significant. Train and test multimodal ML models to predict interview outcomes using both language and prosodic features. Use a filter feature selection method to identify the m prosodic features that are the most relevant to each outcome. Repeat question **(c)** using the prosodic features only. Following that, combine the prosodic features with the language features to create a multimodal feature set, and train a model using this combined data. Discuss how each modality contributes to the overall performance of the model and interpret which features seem to have the most significant impact on predicting interview outcomes.

Note: The prosodic features have been extracted for each response of the interview. You can average those features across all interview responses in order to obtain a single prosodic feature vector per inteview. You can find more information on the prosodic features at [5], which is file 'naim-fg15.pdf'.

(e) (2 points) Explainable ML. Use an explainable ML algorithm for interpreting the decisions and decision-making process of the two types of ML models that you developed in question **(c)**. Discuss the pros and cons of each ML algorithm and the corresponding types of explanations that are provided. You can evaluate the provided explanations in terms of various criteria such as comprehensibility, relevance, and scalability.

Note: You can try different algorithms, such as the EBM-Explainable Boosting Machine (`https://interpret.ml/docs/ebm.html`), the SHAP-SHapley Additive exPlanations (`https://shap.readthedocs.io/en/latest/`), LIME-Local Interpretable Model-Agnostic Explanations (`https://github.com/marcotcr/lime`), etc. Please see [1], uploaded under filename 'explainability.pdf' for additional discussion and resources on ML explainability.

(f) (Bonus, 2 points) Experimenting with transformers. Use a pre-trained transformer-based model (e.g., quantized Llama, minGPT) to estimate the interview outcomes based on the provided transcripts. Experiment with prompt engineering or task-specific prompts to guide the model in adapting to each classification objective, including incorporating a few labeled example transcripts within the prompt. In addition to estimating the interview outcome, also prompt the model to generate an textual explanation about its decision. Please use the same

evaluation metrics as in **(d)** and similar evaluation criteria for the explanation as in **(e)**. Please provide a discussion on how this model compare to the previous models.

Note: You can use the following github repo: `https://github.com/karpathy/minGPT`

(g) (2 points) Presentation. Create a presentation of your work. The presentation will provide the main gist of your work, including the problem statement, your methodology, and the main results from your experiments. **Add visuals so that people understand the main concepts.** Each team will have 4 minutes to present followed by 4 minutes of questions.

Note: Each team will present in class on **December 2, 4, and 6** during class time (**10.10-11pm MT**). The assigned time slot for your team has been announced on CANVAS. **Participation in the presentation for all three days (December 2, 4, 6) is mandatory will count toward the in-class participation grade. Members from other teams will be called on by name to ask questions during the presentations.**

(h) (2 points) Teamwork. Please follow the survey link below, in which you will need to discuss the level of engagement of your team members. The feedback that your team members will provide for one another will be used as an indication of each member's contribution to this group assignment.

Survey Link: `https://forms.gle/P2GuKRVmoUqfZXL76`

References

[1] V. Belle and I. Papantonis. Principles and practice of explainable machine learning. *Frontiers in big Data*, 4:688969, 2021.

[2] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128, 2014.

[3] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706, 2013.

[4] C. Leaper and R. D. Robnett. Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. *Psychology of women quarterly*, 35(1):129–142, 2011.

[5] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–6. IEEE, 2015.

[6] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236, 2008.

[7] K. B. Tølbøll. Linguistic features in depression: A meta-analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(2):39, 2019.

[8] R. Verrap, E. Nirjhar, A. Nenkova, and T. Chaspari. 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, 2022.