CSCI 5622
Homework 4: Designing ML models for real-world problems
Draft assignment for the purpose of picking projects; instructions are subject to minor changes

---

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. We tried to take survey preferences into account, but they are not guaranteed. Please see assignments on CANVAS.

3. You will present your work with your team in class on **April 21 or April 23** during class time (**3.35-4.50pm 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. **Each team will have roughly 6 minutes to present, followed by roughly 4 minutes of questions.**

4. The final report is due on **April 25, 2025 @ 5.00pm**. 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 **16 points** (out of 100 total for the class).

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 1: Mitigating socio-demographic bias in language-based machine learning models of depression

Patients with depression depict differences in language use compared to their healthy counterparts, such as an increased use of first-person singular pronoun use, increased use of negative emotion words, and decreased use of positive emotion words [7]. Yet, the use of language also carries evidence on the speaker's socio-demographic characteristics, such as gender, race, and ethnicity. For example, in a study that analyzed English text, women used more words related to psychological and social processes, while men referred more to object properties and impersonal topics [6]. A meta-review further found that women were more likely to use more tentative language than men [4]. This study explores the effectiveness of language-based ML models for automatically estimating the degree of depression severity, as well as gender and race/ethnicity bias in these models.

The data for this study comes from the Extended Distress Analysis Interview Corpus (E-DAIC) dataset [2], which contains clinical interviews designed to support the diagnosis of psychological distress conditions such as depression. More details about the dataset, including the experimental setup and type of data, can be found here: `https://dcapswoz.ict.usc.edu/`. The data uploaded on CANVAS includes interview transcripts from 190 participants. The 'DAIC demographic data.xlsx' file contains three tabs with information on participants' gender, race/ethnicity, and depression severity (PHQ_Score). The folder 'E-DAIC_Transcripts' includes csv files (one per participant) with the transcript data. Each filename, named as 'x_Transcript.csv', where x is the participant ID, includes the transcripts from the responses provided by participant x during the clinical interview. In the following analysis, participants will be grouped by gender, including female and male participants, and by race/ethnicity, including African American, Hispanic, and White American participants, along with the intersections of these categories. **Due to the small sample size of the other groups, the remaining participants will be excluded from this analysis.**

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:

- 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) Classifying for gender. Use one tree-based ML model of your choice and one deep learning ML model of your choice to classify participants in terms of gender. Explore a filter feature selection method of your choice to identify the n features that are the most informative of gender based on the provided data. Experiment with different values of n. Please report the simple classification accuracy A and balanced classification accuracy BA.

Note: The simple classification accuracy A and balanced classification accuracy BA are defined as follows:
$A = \frac{\#\text{correctly classified samples}}{\text{total \# samples}}$
$BA = 0.5 \cdot \frac{\#\text{correctly classified samples for depression}}{\text{total \# samples for depression}} + 0.5 \cdot \frac{\#\text{correctly classified samples for no depression}}{\text{total \# samples for no depression}}$

(c) (2 points) Classifying for race/ethnicity. Use one tree-based ML model of your choice and one deep learning ML model of your choice to classify participants in terms of race/ethnicity. Explore a filter feature selection method of your choice to identify the m features that are the most informative of race/ethnicity based on the provided data. Experiment with different values of m. Please report the simple classification accuracy A and balanced classification accuracy BA.

(d) (2 points) Estimating depression severity. Use one tree-based ML and one deep learning ML algorithm of your choice to estimate the degree of depression for each participant. Explore a filter feature selection method of your choice to identify the k features that are the most informative of depression based on the data. Please report the Pearson's correlation r and absolute relative error (RE) between the estimated and actual PHQ-8 scores for all participants included, as well as separately by gender (i.e., female, male), race/ethnicity (African American, White American, Hispanic), and their intersection (African American female, White American female, etc.). Experiment with different values of k. Please discuss your findings (e.g., Is there any overlap between the features that are the most informative of demographic characteristics and the ones that are the most informative of depression severity? Are there differences in performance among the different demographic groups based on gender, race/ethnicity, and their intersection? For which demographic groups (gender×race/ethnicity combination) does the ML model depict the worst performance?).

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

(e) (2 points) Mitigating bias via reducing socio-demographic dependencies in features. Remove the n most informative features of gender and the m most informative features of race/ethnicity from the original feature set. Use the updated feature set and the same ML models as in the previous questions to conduct depression severity estimation. In addition to the aforementioned feature selection, please use one more in-processing method to mitigate potential bias. Please report the results similar to **(d)** and discuss your findings.

Note: You can use the following toolbox for in-processing methods for de-biasing:
`https://github.com/ahxt/fair_fairness_benchmark`
Please discuss the results of the revised ML models grounded in your findings from question

(d), after identifying the demographic group for which the ML model performed the worst. Depending on the value of n and m, you can use the combination of features that are most informative to gender and most informative to race/ethnicity, or the features that are common between the two.

(f) (2 points) Experimenting with transformers. Use a pre-trained transformer-based model (e.g., quantized Llama, minGPT) to conduct depression severity estimation 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. Please use the same evaluation metrics as in **(d)**. How does the performance of this model compare to the previous models? Please provide a discussion for all participants and for different participant groups.

Note: You can find a useful github repo here: `https://github.com/karpathy/minGPT`

(g) (4 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 **April 21 or April 23** during class time (**3.35-4.50pm MT**). The assigned time slot for your team has been announced on CANVAS.

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) (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) (4 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 **April 21 or April 23** during class time (**3.35-4.50pm MT**). The assigned time slot for your team has been announced on CANVAS.

Study 3: Privacy-enhancing language-based machine learning for detecting effective language use in job interviews

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 a candidate's effective communication strategies to identify strengths and areas for improvement.

This study will design ML models to detect ineffective communication in job interviews, and particularly, the degree of explanation of interview responses. It will classify an interview response among four classes (i.e., under-explained, succinct, comprehensive, over-explained; see Table 1). Additionally, the study will analyze how language features depend on the speaker's identity and work to minimize these potential dependencies within the ML models, offering privacy-enhancing ML approaches.

The data for this study come from the VetTrain dataset [8] with a total of 38 participants. The folder 'VetTrain_Transcripts' includes the interview transcripts. Each filename, named as 'Px_transcript.csv', where x is the participant ID, includes the transcripts of the interview from participant x. Each file has four columns: the utterance type (see file 'README_transcript.txt' for more information), the start time of the utterance, the end time of the utterance, and the transcript of the utterance. 'Q<id>' represents the question asked by the interviewer, starting from Q1, ending at Qn, where n is the total questions asked. 'A<id>' represents the answer by the veteran to the corresponding question, starting from A1, ending at An, where n is the total questions asked. The file 'Behavioral Annotation Codes.csv' has three columns, including the participant ID, the question ID, and the label of the response (i.e., under-explained, succinct, comprehensive, over-explained).

In the following, we will assume that all utterances from both the interviewer and the interviewee that correspond to a single question/response pair (i.e., 'Q<id>' and 'A<id>') form a single sample.

Class	**Definition**	N
Under-explained	A response that is short but does not fully answer the interviewer's question, and often ends abruptly.	23
Succinct	A concise/short/brief/to-the-point response that fully answers the interviewers' question.	107
Comprehensive	A detailed but to-the-point response that answers the interviewers' questions fully and in a detailed manner.	122
Over-explained	A very long response with excess verbiage. It often goes into too much detail, which potentially affects the coherence of the answer. It can also include repetitive information.	34

Table 1: Definitions and the number of samples (N) per class for degree of explanation.

(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) Classifying for speaker identity. Use one tree-based ML model of your choice and one deep learning ML model of your choice to classify participants in terms of speaker identity. Explore a filter feature selection method of your choice to identify the n features that are the most informative of speaker identity based on the provided data. Experiment with different values of n. **Randomly split the data into 5 folds, so that samples from one participant can be found both in the train and test sets.** Please report the simple classification accuracy A and balanced classification accuracy BA.

Note: The simple classification accuracy A and balanced classification accuracy BA are defined as follows:

$$A = \frac{\#\text{correctly classified samples}}{\text{total } \# \text{ samples}}$$

$$BA = 0.5 \cdot \frac{\#\text{correctly classified samples for depression}}{\text{total } \# \text{ samples for depression}} + 0.5 \cdot \frac{\#\text{correctly classified samples for no depression}}{\text{total } \# \text{ samples for no depression}}$$

(c) (2 points) Classifying between under-explained and succinct. Use one tree-based ML and one deep learning ML algorithm of your choice to conduct a binary classification task classifying between under-explained and succinct. Explore a filter feature selection method of your choice to identify the k features that are the most informative of the task based on the data. **Split the data into 5 participant-independent folds, so that samples from one participant can be found either in the train set or in the test set.** Use only the responses that correspond to the two classes of interest. Please report the simple classification accuracy A, balanced classification accuracy BA. Experiment with different values of k. Please discuss your findings (e.g., Is there overlap between the features that are the most informative of speaker identity and the ones that are the most informative for under-explained vs succinct explanation? Are there differences in performance of the classifier among the different speakers?)

Note: You can either run a 4-way classification task, or three binary classification tasks (i.e., under-explained and succinct vs comprehensive and over-explained; followed by under-explained vs succinct; and comprehensive vs over-explained).

(d) (2 points) Classifying between comprehensive and over-explained. Repeat question (c) for the binary classification task between comprehensive and over-explained.

(e) (2 points) Reducing speaker identify dependencies in features. Remove the n most informative features of speaker from the original feature set. Using the updated feature set and the same ML models as in questions **(c)** and **(d)**, repeat the two binary classification tasks for the degree of explanation. In addition to the aforementioned feature selection, please use one more in-processing method to mitigate the potential effect of speaker identity in the data. Please report the results similar to **(c)** and **(d)** and discuss your findings.

Note: You can use the following toolbox for in-processing methods for reducing the effect of speaker identity on the data:

`https://github.com/ahxt/fair_fairness_benchmark`

(f) (2 points) Experimenting with transformers. Use a pre-trained transformer-based model (e.g., quantized Llama, minGPT) to estimate the degree of explanation 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 degree of explanation for a response, also prompt the model to generate a textual reasoning about its decision. Please use the same evaluation metrics as in **(c)** and **(d)**. 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) (4 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 **April 21 or April 23** during class time (**3.35-4.50pm MT**). The assigned time slot for your team has been announced on CANVAS.

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.