

# Data Augmentation for Reliability and Fairness in Counselling Quality Classification

Vivek Kumar<sup>1,3</sup><sup>a</sup>, Simone Balloccu<sup>2,3</sup><sup>b</sup>, Zixiu Wu<sup>1,3</sup><sup>c</sup>,  
Ehud Reiter<sup>2</sup><sup>d</sup>, Rim Helaoui<sup>3</sup><sup>e</sup>, Diego Reforgiato Recupero<sup>1</sup><sup>f</sup> and Daniele Riboni<sup>1</sup><sup>g</sup>

<sup>1</sup>University of Cagliari, Cagliari, Italy

<sup>2</sup>University of Aberdeen, Aberdeen, U.K.

<sup>3</sup>Philips Research, Eindhoven, The Netherlands

**Keywords:** AI Fairness, Motivational Interviewing, Counselling, Dialogue, Natural Language Processing, Machine Learning.


**Abstract:** The mental health domain poses serious challenges to the validity of existing Natural Language Processing (NLP) approaches. Scarce and unbalanced data limits models' reliability and fairness, therefore hampering real-world application. In this work, we address these challenges by using our recently released Anno-MI dataset, containing professionally annotated transcriptions in motivational interviewing (MI). To do so, we inspect the effects of data augmentation on classical machine (CML) and deep learning (DL) approaches for counselling quality classification. First, we adopt augmentation to balance the target label in order to improve the classifiers' reliability. Next, we conduct the bias and fairness analysis by choosing the therapy topic as the sensitive variable. Finally, we implement a fairness-aware augmentation technique, showing how topic-wise bias can be mitigated by augmenting the target label with respect to the sensitive variable. Our work is the first step towards increasing reliability and reducing the bias of classification models, as well as dealing with data scarcity and imbalance in mental health.


## 1 INTRODUCTION


Recent advancements in Natural Language Processing (NLP) captured the interest of research community in healthcare (Kumar et al., 2020b; Dessì et al., 2020; Locke et al., 2021; Kumar et al., 2020a), including mental health and its subdomains such as depression, anxiety or substance abuse (Le Glaz et al., 2021). However, real world application of clinical NLP is hampered by multiple elements such as domain complexity, rigorous accuracy and reliability standards and data scarcity (Ibrahim et al., 2021). Lastly, recent research highlighted critical concerns on artificial intelligence (AI) fairness (Chouldechova


and Roth, 2020; John-Mathews et al., 2022), that is imperative to address when applying NLP to mental health.


As the first step towards addressing these issues, we adopt data augmentation to improve AI reliability and fairness in the context of scarce mental health data. We leverage our recently released dataset Anno-MI (Wu et al., 2022), consisting of professionally annotated therapy transcriptions in MI (Miller and Rollnick, 2012; Rollnick et al., 2008). We model a classification task, targeting overall therapy quality, one of Anno-MI most unbalanced labels, using each therapist's utterance as input data. In the fairness context, we inspect therapy topics, e.g., "smoking cessation", "reducing alcohol consumption" or "diabetes management" as the sensitive variable. We conduct a quantitative analysis of the effects of data augmentation to balance target and sensitive variables. Our experimental results show little to no effect on Classical Machine learning (CML) classifiers, but prove that Deep Learning (DL) ones benefit from augmented data, showing consistent improvement in both accu-


<sup>a</sup>  <https://orcid.org/0000-0003-3958-4704>

<sup>b</sup>  <https://orcid.org/0000-0002-9812-5092>

<sup>c</sup>  <https://orcid.org/0000-0002-3679-5701>

<sup>d</sup>  <https://orcid.org/0000-0002-7548-9504>

<sup>e</sup>  <https://orcid.org/0000-0001-6915-8920>

<sup>f</sup>  <https://orcid.org/0000-0001-8646-6183>


<sup>g</sup>  <https://orcid.org/0000-0002-0695-2040>

Table 1: The overall distribution of high and low quality therapy utterances.

Dataset	Total utterances (no.)	High quality (%)	Low quality (%)
Anno-MI	2601	91%	9%
Anno-AugMI	5302	45%	55%
Anno-FairMI	9154	50%	50%

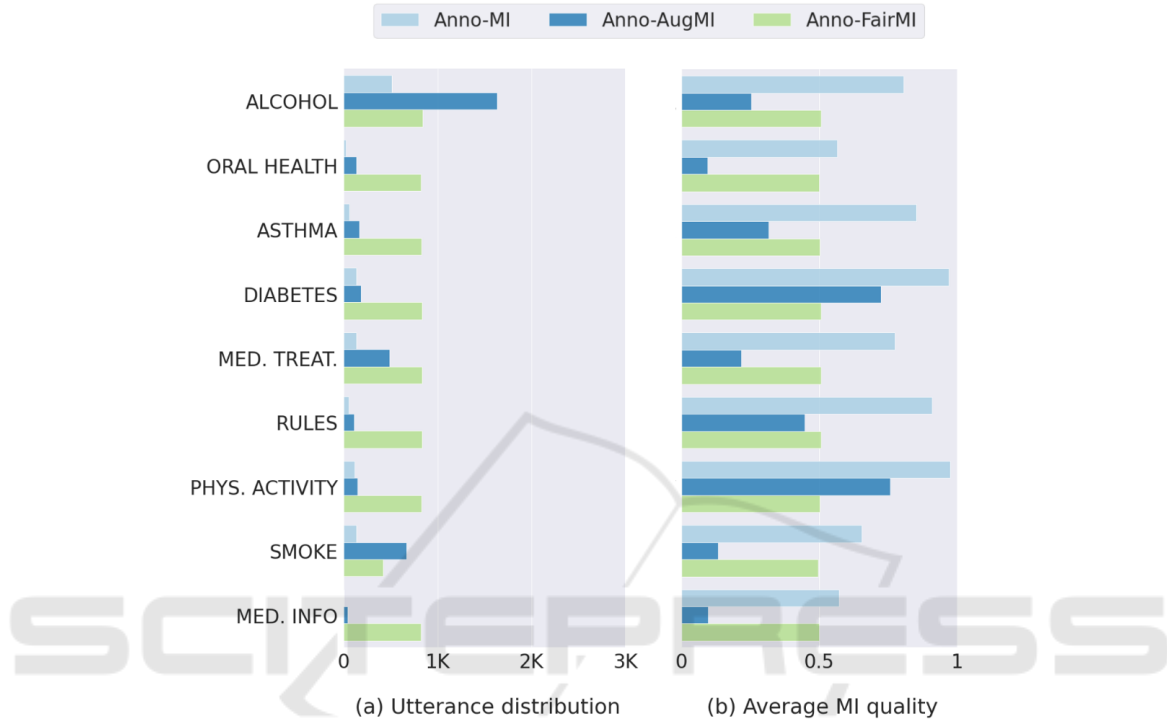


Figure 1: Sensitive variable statistics for each dataset. We show topic-wise (a) utterances distribution and (b) average therapy quality. For brevity, only common topics for each dataset are shown.

racy and reliability. Fairness assessment shows that more work on augmentation is required to properly mitigate eventual classification BIAS.

## 2 MATERIAL AND METHODS

Anno-MI<sup>1</sup> (Wu et al., 2022) contains 110 high-quality and 23 low-quality MI conversational dialogues from a total of 44 topics e.g.: “smoking cessation”, “diabetes management”, “anxiety management” and others. Therapy quality indicates the therapist’s adherence to “general counseling principles taken from the literature on client-centered counseling” (Pérez-Rosas et al., 2019). Therapy quality distribution in Anno-MI is heavily skewed towards high-quality (HQ-MI) utterances. This is because the conversations that constitute the dataset belong to MI training videos, which rarely showcase low-quality (LQ-MI)

<sup>1</sup>Data available at <https://github.com/uccollab/AnnoMI>

counseling scenarios. We employ data augmentation to overcome these issues.

We leverage NL-Augmenter<sup>2</sup> (Dhole et al., 2021) to develop a 11-step augmentation pipeline, each one taking one utterance as input. Therefore, for each given utterance, we obtain  $n \geq 11$  augmentations (due to certain augmenters potentially producing multiple alternatives for the same utterance). The adopted augmentation techniques include noising, paraphrasing and sampling (Li et al., 2022). Since our augmentation process is unsupervised, we avoid using techniques that could lead to semantic changes with respect to the original utterance. With this setup, we generate two augmented versions of Anno-MI, targeting classifier reliability and fairness, respectively.

<sup>2</sup>Code available at <https://github.com/GEM-benchmark/NL-Augmenter>

## 2.1 Problem Statement

We model a binary classification task to detect therapy quality from a single therapist utterance. We assign each therapist utterance, to the corresponding conversation quality, in order to formulate the positive and negative examples for our task. Indeed, assessing the quality of MI sessions can boost therapist training and skills assessment, as confirmed from the existing related work on empathy modelling (Xiao et al., 2012; Gibson et al., 2015; Gibson et al., 2016; Wu et al., 2020), automatic coding of therapeutic utterances (Atkins et al., 2014; Xiao et al., 2016; Cao et al., 2019) and session-level therapist performance (Flemotomos et al., 2022). Given the previously mentioned quality skewness, the target variable represents the first potential source of classification unreliability. In this context, we introduce Anno-AugMI, consisting of all the therapist utterances from Anno-MI, augmented in order to balance quality proportion. Anno-AugMI creation proceeds in a topic-agnostic fashion, with the goal of obtaining a roughly balanced amount of HQ-MI and LQ-MI utterances across the entire dataset. Since therapy quality is the target of our classifiers, we call this procedure *target-aware augmentation*. No check is in place with regards to which utterances are augmented, meaning that *target-aware augmentation* merely iterates over the dataset and augments every low-quality utterance until the target label is balanced.

To assess classification fairness, it is necessary to identify the sensitive variable and field-test it with the employed classifiers. We choose the therapy topic (MI-topic) as our sensitive variable, as inter-topic fairness guarantees stable performances across a wide range of therapy goals, and because therapy quality in Anno-MI is also unbalanced at topic-level (as shown in Figure 1). To address fairness, we introduce Anno-FairMI, consisting of all the therapist utterances from Anno-MI, augmented to balance therapy quality proportion with respect to MI-topic. Anno-FairMI creation proceeds in a topic-aware fashion, with the goal of having the same amount of HQ-MI and LQ-MI utterances for each MI-topic. Since MI-topic is the sensitive variable of our classifier, we call this procedure *fairness-aware augmentation*. This last procedure introduces the necessity to cut out those MI-topic which have no low-quality example since augmentation would have been impossible. As a result, Anno-MI and Anno-AugMI share all the 44 topics (134 conversations), while Anno-FairMI keeps only 9 topics (55 conversations), resulting in a much lower pre-augmentation data size.

The comparative distribution of topic-wise utterances, and average therapy quality per topic is shown in Figure 1. The overall distribution of labels in Anno-MI, Anno-AugMI and Anno-FairMI is shown in Table 1.

## 3 EXPERIMENTS AND RESULTS

We design a series of experiments, where each experiment's input is based on the output of the preceding ones. The experimental setup is as follows:

- Therapist utterances quality classification of Anno-MI.
- Augmentation of Anno-MI to balance therapy quality.
- Therapist utterances quality classification of Anno-AugMI.
- Fairness assessment of Anno-AugMI.
- Augmentation of Anno-MI based on MI-topic.
- Therapist utterances quality classification of Anno-FairMI.
- Fairness assessment and BIAS mitigation of Anno-FairMI.

We use Support Vector Machine (SVM) and Random Forest (RF) as CML classifiers, and a Bidirectional Long Short Term Memory (Bi-LSTM) with Word2Vec pre-trained word embedding for the embedding layer. We use balanced accuracy and F1 score as performance evaluation metrics for classifiers. We use one universal test set for all the experiments, created by extracting 400 high quality and 100 low quality utterances from Anno-MI. The rest of the data is considered as training set and constitutes the basis for the augmentation.

To assess the fairness and mitigate eventual BIAS of our classifiers we use Microsoft FairLearn<sup>3</sup> (Bird et al., 2020) and inspect Selection Rate (SR), False Negative Rate (FNR) and Balanced Accuracy (BA) as evaluation metrics. Where applicable, we adopt "Threshold Optimization" with BA as the target and False Negative Parity as the fairness constraint. Since Anno-MI and Anno-AugMI contain multiple topics that lack LQ-MI utterances, it is not possible to split training, test and validation data so that each partition contains both therapy quality classes. The presence of degenerate labels prevents BIAS mitigation, so for these datasets we only evaluate the initial metrics values.

The classification results of CML and DL approaches for each of the three datasets are summed up

<sup>3</sup>Code available at <https://github.com/fairlearn/fairlearn>

Table 2: Performance of CML and DL approaches with Anno-MI, Anno-AugMI, Anno-FairMI. For each dataset we report Balanced Accuracy and F1 score calculated with regards to MI quality.

Dataset	SVM		Random Forest		Bi-LSTM (DNN)	
	Bal.Acc.	F-1	Bal.Acc.	F-1	Bal.Acc.	F-1
Anno-MI	50.00	44.44	50.75	46.34	50.00	44.44
Anno-AugMI	48.87	38.12	50.37	45.78	<b>73.12</b>	<b>71.85</b>
Anno-FairMI	53.87	48.15	51.00	50.99	64.13	59.50

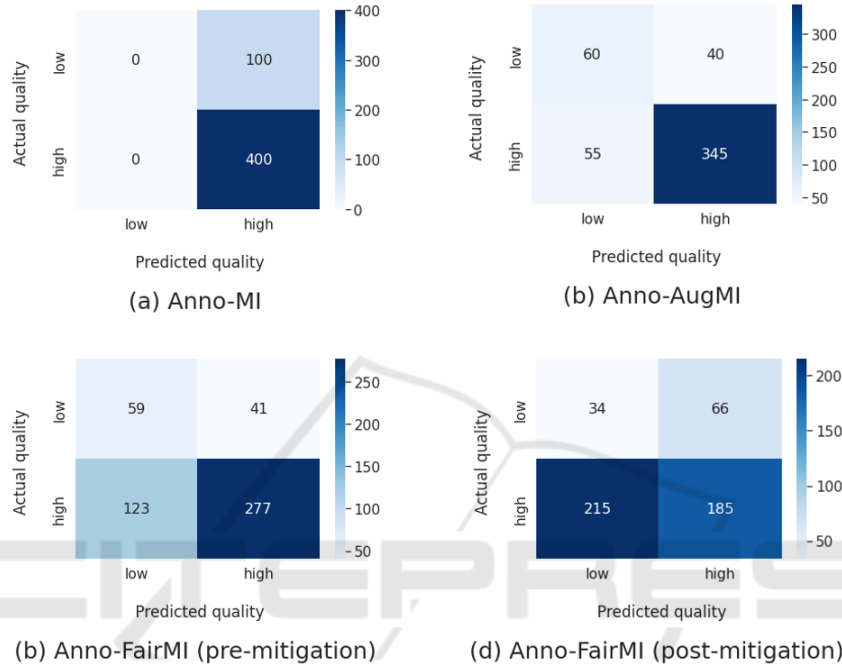


Figure 2: Confusion matrix for the Bi-LSTM trained on each dataset. For Anno-FairMI we provide pre and post-mitigation matrix.

in Table 2. The obtained results are indicative of consistent low performance of the CML with Anno-MI. Our augmentation techniques are quite simple so they do not add prominent features to Anno-MI, which can be very helpful in distinguishing classes with bag-of-words representation. This explains the minor performance improvement of the CML algorithms. Since both SVM and RF did not benefit from data augmentation and are comparable to random classifiers, we do not go any further with their analysis. On the other hand, Bi-LSTM model shows significant performance enhancement of 23-14% for Anno-AugMI and Anno-FairMI respectively over Anno-MI. Further considerations can be drawn by looking at the confusion matrix in Figure 2. The initial model, trained on Anno-MI, suffers from the skewed therapy quality distribution and is unable to recognise LQ-MI utterances. This problem also reflects on HQ-MI, with no false positives at all. With *target-aware augmentation* on Anno-AugMI we see more promising results with about 40% of false positives and 14% of false

negatives. Finally, with *fairness-aware augmentation* on Anno-FairMI we see pretty much no change in LQ-MI classification, but a considerable drop with HQ-MI, with about 30% false negatives. This can be motivated by the reduced amount of topics in Anno-FairMI, making the Bi-LSTM suffer from the unseen ones in test set. In both cases, data augmentation led to an accuracy improvement, which makes our approach promising for future developments (Rice and Harris, 2005).

Fairness metrics values for each dataset are showed in Figure 3. SR and FNR are apparently ideal for Anno-MI, but this is purely related to the low BA value. Anno-AugMI shows more unbalanced values for SR and FNR, but higher BA than Anno-MI across pretty much every topic. For Anno-FairMI, BIAS mitigation can be ran because of the absence of degenerate labels in training set. Pre-mitigation, Anno-FairMI shows generally more balanced SR, lower FNR and higher BA than the other two datasets for known topics, and little to no effect after mitiga-

Table 3: The effects of BIAS mitigation on Bi-LSTM trained on Anno-FairMI. For each metric, we report the mean value calculated with regards to the sensitive variable (therapy topic). “TO” stands for “Threshold Optimisation”.

Dataset	Selection Rate	False Negative Rate	Bal. Acc.
Anno-FairMI	67.29	23.94	75.72
Anno-FairMI + TO	19.60	72.86	21.89

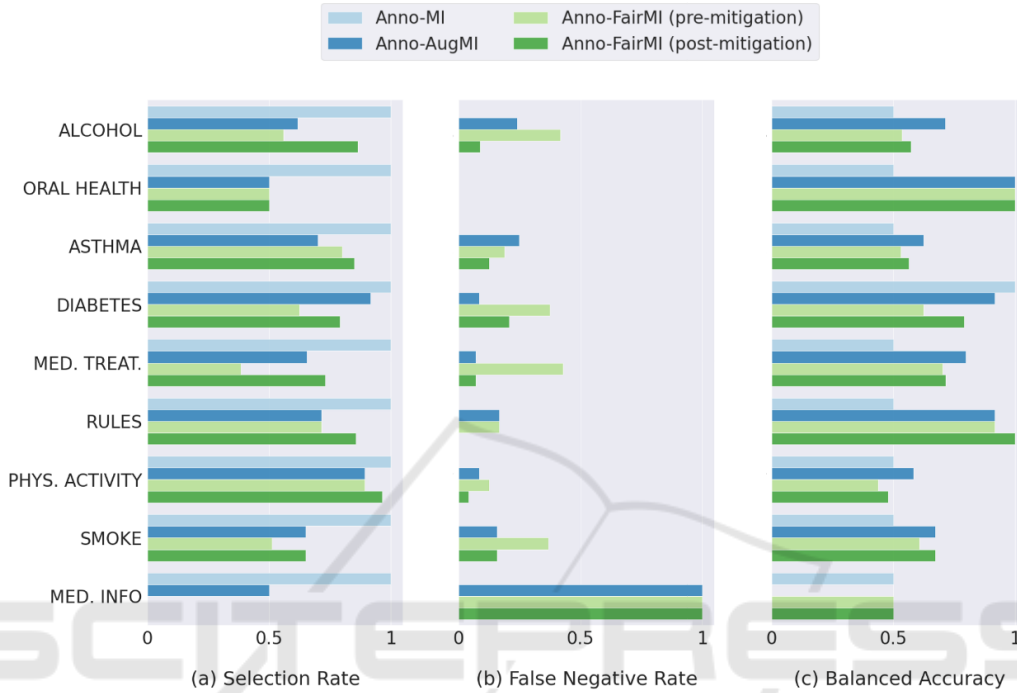


Figure 3: Fairness assessment and BIAS mitigation for Bi-LSTM on each dataset. For brevity, only common topics for each dataset are shown.

tion. However, moving to unseen topics the overall Bi-LSTM performances greatly worsened, with compromised classification (Figure 2) and fairness metrics dropping significantly (Table 3).

## 4 CONCLUSION AND FUTURE WORK

In this work we employed data augmentation to balance target and sensitive variable on our dataset of MI transcriptions Anno-MI, resulting in two augmented datasets, namely Anno-AugMI and Anno-FairMI. We evaluated our approaches on a classification task, aimed at recognising therapy quality. Our results show a promising accuracy increase for DL classifiers by using augmented datasets, especially Anno-AugMI. This motivates us to consider other target attributes in future works, such as client talk type or therapist behaviour, also extending to other tasks like forecasting. The fairness assessment and BIAS mitigation show

that Anno-FairMI is too sensitive to unseen topics, opening interesting future work on the adoption of more advanced augmentation techniques. Overall, we consider *target-aware augmentation* effective at addressing the challenges of unbalanced and scarce data in the mental health domain. Finally, we aim to perform human evaluation of the developed classifier, to sanity check the reliability of the obtained results.

## ACKNOWLEDGEMENTS

This work is supported by the EU’s Marie Curie training network PhilHumans—Personal Health Interfaces Leveraging Human–Machine Natural Interactions under Agreement 812882.

## REFERENCES

Atkins, D. C., Steyvers, M., Imel, Z. E., and Smyth, P. (2014). Scaling up the evaluation of psychotherapy:



- evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11.
- Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.
- Cao, J., Tanana, M., Imel, Z., Poitras, E., Atkins, D., and Srikumar, V. (2019). Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611.
- Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89.
- Dessi, D., Helaoui, R., Kumar, V., Recupero, D. R., and Riboni, D. (2020). TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study. In Consoli, S., Recupero, D. R., and Riboni, D., editors, *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020*, volume 2596 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org.
- Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahendiran, A., Mille, S., Srivastava, A., Tan, S., et al. (2021). Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.
- Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., et al. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, 54(2):690–711.
- Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., and Narayanan, S. S. (2016). A Deep Learning Approach to Modeling Empathy in Addiction Counseling. In *Proc. Interspeech 2016*, pages 1447–1451.
- Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., and Narayanan, S. S. (2015). Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.
- Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D., and Denniston, A. K. (2021). Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*, 3(4):e260–e265.
- John-Mathews, J.-M., Cardon, D., and Balagué, C. (2022). From reality to world. a critical perspective on ai fairness. *Journal of Business Ethics*, pages 1–15.
- Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N., and Verma, A. (2020a). Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In *Advances in data science and management*, pages 435–442. Springer.
- Kumar, V., Recupero, D. R., Riboni, D., and Helaoui, R. (2020b). Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. *IEEE Access*, 9:7107–7126.
- Le Glaz, A., Haralambous, Y., Kim-Dufoir, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5):e15708.
- Li, B., Hou, Y., and Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*.
- Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., and Kitchen, G. B. (2021). Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, 38:4–9.
- Miller, W. R. and Rollnick, S. (2012). *Motivational interviewing: Helping people change*. Guilford press.
- Pérez-Rosas, V., Wu, X., Resnicow, K., and Mihalcea, R. (2019). What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935. Florence, Italy. Association for Computational Linguistics.
- Rice, M. E. and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: Roc area, cohen’s d, and r. *Law and human behavior*, 29(5):615–620.
- Rollnick, S., Miller, W. R., and Butler, C. (2008). *Motivational interviewing in health care: helping patients change behavior*. Guilford Press.
- Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E., Recupero, D. R., and Riboni, D. (2022). Anomi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.
- Wu, Z., Helaoui, R., Kumar, V., Reforgiato Recupero, D., and Riboni, D. (2020). Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 497–502.
- Xiao, B., Can, D., Georgiou, P. G., Atkins, D., and Narayanan, S. S. (2012). Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE.
- Xiao, B., Can, D., Gibson, J., Imel, Z. E., Atkins, D. C., Georgiou, P. G., and Narayanan, S. S. (2016). Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Interspeech*, pages 908–912.