Designing AI Systems to Support Human Writing

Yuqing Wu
Advised by: Qian Yang, Tony Wang

Project 1: Human AI Interactions to Assist Identity Reflection in Writing

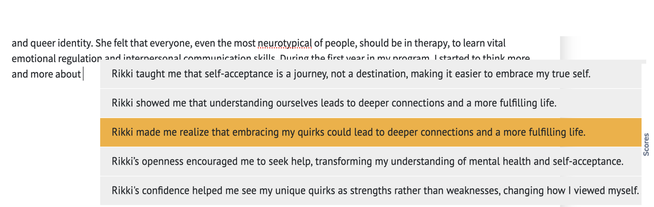
The Study

Study Structure
The study is composed of two phases:

- Prewriting: describe experiences and relationship with community that influenced your identity, without AI.
- Reflection: reflect on what was written in the prewriting, with AI.

AI Interaction Feature
The participant can ask AI to autocomplete their sentences or use $...$ to send request to AI. The prompts and their writing are then sent to GPT-4o-mini and the response is displayed to them.

We adapted CoAuthor Toolkit [1] to build our study platform.



Measures

Agency: the degree to which someone is able to make change in their lives or influence others.

Communion: the degree to which someone demonstrate or experience interpersonal connection.

Agency and communion are two important measures of clear sense of identity in writing [2].

We measured agency and communion by calculating the semantic similarity between written text and two dictionaries that contains keywords representing agency and communion [3].

Externality: we used the external score to characterize the writing process. Higher external scores represent more abstract statements while lower external scores represent episodic reexperiencing where they describe concrete details about the situation as if they were reexperiencing the event.

We measured externality of writing using a distilBERT Huggingface model [4].

Key Cognitive Interactions

A dip in the external graph generally represents the participant writing about detailed descriptions of some life experience and then coming back to reflect at a more abstract level. Observe that these dips correspond to sharp increases in the agency and communion graphs.

Next steps:

- Explore more interesting cognitive interactions:
  - AI reminded humans of keywords they are interested in writing about, sometimes leading to deep reflection.
  - Human and AI coming up with ideas and having a conversation.
- Adding interventions to encourage positive interactions and alert negative interactions in studies.

References

[1] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 388, 1–19. https://doi.org/10.1145/3491102.3502030
[2] McAdams, D. P., & McLean, K. C. (2013). Narrative Identity. *Current Directions in Psychological Science, 22*(3), 233-238. https://doi.org/10.1177/0963721413475622
[3] Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology, 49*(5), 871–887. https://doi.org/10.1002/ejsp.2561
[4] van Genugten, R.D., Schacter, D.L. Automated scoring of the autobiographical interview with natural language processing. *Behav Res* **56**, 2243–2259 (2024). https://doi.org/10.3758/s13428-023-02145-x

Project 2: Designing Reward and Simulation For Contextual Bandit Prompt Recommendation System

The Problem

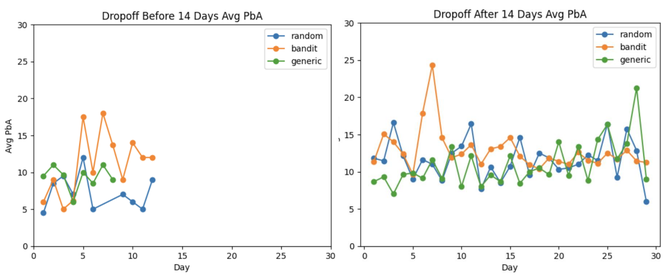
Therapists give writing prompts to patients to journal to improve their mental health. Our goal is to build a recommender system that recommends writing prompts to patients such that they continue to journal.

We collected a dataset in a longitudinal study where we ask people to journal every day given some prompts. The dataset includes the following information:

- Patient's prior mental health
- Patient's prior journaling experience
- Keystroke logs for each journal written (which contains the time it was written as well)
- Prompt used
- Category of each prompt

Designing Reward

We used a Point-Based Engagement Measurement Algorithm (PbA) to compute how much effort a writer puts in during the writing process [1]. The PbA algorithm creates writing bursts based on the pause time between keystrokes. We normalized the pause time for separating bursts by a factor of the person's median pause time [2].



We have a dataset where people journal based on prompts that were chosen randomly, using a bandit algorithm that optimize on wordcount, or a static generic prompt. PbA increases for all engaged individuals at the end.

Simulation With Oracle

To simplify the problem, we worked on prompt categories instead of individual prompts.

Because there is no ground truth on what prompt should be given in our dataset, we want to create an oracle that approximates the ground truth.

Reward Model: we fit the NegativeBinomial regression [3] on our dataset.

Oracle: we brute force each action and find the action that maximizes the reward computed with our reward model.

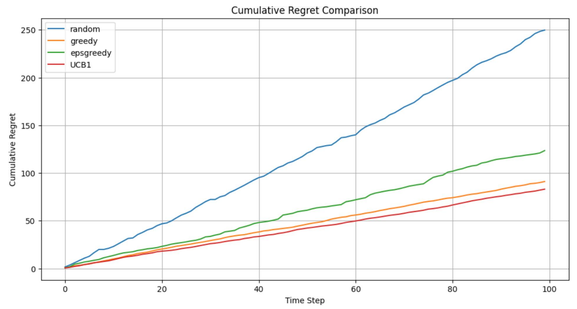
Contextual Bandits: we used different bandit models from MABWiser[4] to predict prompt categories and train these models with the oracle reward model.

Simulation: we generate two features (prior mental health and journaling experience) randomly and use those as contexts to the bandit models.

Regret: we then calculate the cumulative regret by summing regret at each timestep, which is defined as

Regret = oracle reward – contextual bandits reward

We now compare how fast different bandit algorithms learn with the oracle.



Next steps:

- Deploying the bandit algorithms in user studies to see if there's a difference in prompt recommendation when the reward is PbA.
- Experimenting with actual prompts instead of prompt categories and trying to put the prompts in embedding space.

References

[1] M. Liu, R. A. Calvo, A. Pardo and A. Martin, "Measuring and Visualizing Students' Behavioral Engagement in Writing Activities," in IEEE Transactions on Learning Technologies, vol. 8, no. 2, pp. 215-224, 1 April-June 2015, doi: 10.1109/TLT.2014.2378786.
[2] Taisa Kushner and Amit Sharma. 2020. Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums. In Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 2906–2912. https://doi.org/10.1145/3366423.3380056
[3] Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychol Bull. 1995 Nov;118(3):392-404. doi: 10.1037/0033-2909.118.3.392. PMID: 7501743.
[4] E. Strong, B. Kleynhans and S. Kadioğlu, "MABWiser: A Parallelizable Contextual Multi-Armed Bandit Library for Python," *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, 2019, pp. 909-914, doi: 10.1109/ICTAI.2019.00129.