AI-Curated Democratic Discourse

Jason Eisner, Daniel Khashabi, Ziang Xiao, Andrew Perrin {eisner,danielk,ziang.xiao,aperrin}@jhu.edu

We seek ways to make the social media experience more prosocial. We will develop a new user interface designed to increase the rate of substantive and constructive conversations, including conversations across political differences and conversations between like-minded strangers.

Specifically, we will use generative AI to:

1. augment the current conversation by showing relevant high-quality posts from other conversations;
2. react to a user's draft post with advice and simulated replies while they are still writing it.

This design goes beyond the traditional threaded conversation model (Usenet, Reddit, Facebook, Twitter, Nextdoor) where disjoint conversations grow one post at a time. It situates posts in a broader curated landscape of viewpoints and supporting information. Users have varying reasons to use social media, but we conjecture that they will sometimes look at good argumentation and competing viewpoints if we make these easy and enjoyable to see.

User interfaces shape user behavior. In this more diverse landscape, posters will have to raise their game. They will be challenged more often by direct responses from strangers, or indirectly by the automatic display of related posts alongside theirs. Thus it becomes harder for them to get away with lazy or specious arguments. We will help them as they write their posts, by previewing simulated reactions and offering suggestions before they submit.

How Can AI Help?

We will rely on the remarkable ability of recent large language models like GPT-4 to closely read human posts in context. LLMs have already been used for clustering text [7, 8], describing differences between different types of text [9], evaluating conversational text on multiple subjective criteria [1, 3], and rewriting text [6].

We anticipate using LLMs for tasks such as these, where P is a post or draft post:

- Scoring P on dimensions such as quality, enjoyability, relevance, non-redundancy, representativeness.
- Classifying the kinds of contributions that P makes to the discussion.
- Finding posts that are similar to P, in that reading them makes it unnecessary to read P.
- Finding posts Q that are relevant to P, in that they provide additional supporting or rebutting arguments, or offer complementary perspectives.
- Lightly rewriting a relevant post Q (if needed) so that it is a more suitable reply to P.
- Synthesizing typical replies and reactions to a draft post P.
- Generating advice about how to improve a draft post P.
- Generating moderation posts that intervene in a conversation.
- Extracting common topics, values, claims, and presuppositions from collections of posts and organizing these by relatedness, to promote a journalistic understanding of the range of viewpoints.

Better Reading

We will aim to promote high-quality posts—which makes the site more interesting for users—and to puncture the filter bubble. We propose to display posts in a threaded interface that *also* gives access to related posts from other threads. If a post P is on a public matter, we display it at the top of a *stack* of similar posts. Clicking on the stack lets the user explore it in a way that favors showing diverse high-quality posts. The user can also specifically explore replies of different types (agreements, disagreements, evidence-based replies, jokes, personal experiences, predictions, proposals, statements of values, etc.).

Also, alongside the ordinary replies to P, we will show related replies to similar posts from across the site. We will make it easy for users to quote-tweet these related replies when writing their own reply. This cross-fertilizes discussions, makes it easy for users to bring in supporting arguments and counterarguments, and incentivizes high-quality posts.

Better Posting

As a user drafts a post or reply P, we will point them to existing conversations on the topic, which they might prefer to join. As they write, we will display P *as if it had already been posted*, simulating the expected reactions and replies. We will also respectfully suggest how to improve the post so that the site will recommend it.

Resources

To create an initial site before the start of the workshop, we will start with an existing codebase such as the opensource-socialnetwork codebase on GitHub. To populate the site, we will either scrape existing content, or generate it by simulating users [5, 4, 2] prompted by political topics on structured debate sites such as Kialo, DebateWise, and Pol.is.

[Update: We now plan instead to build the site as a Mastodon server, with access to a subset of live Mastodon content.]

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *Computing Research Repository*, arXiv:2212.08073, 2022. Available from: https://arxiv.org/abs/2212.08073.

[2] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *Computing Research Repository*, arXiv:2307.14984, 2023. Available from: http://arxiv.org/abs/2307.14984.

[3] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. Watch your language: Large language models and content moderation. *Computing Research Repository*, arXiv:2309.14517, 2023. Available from: http://arxiv.org/abs/2309.14517.

[4] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Computing Research Repository*, arXiv:2304.03442, 2023. Available from: http://arxiv.org/abs/2304.03442.

[5] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. *Computing Research Repository*, arXiv:2208.04024, 2022. Available from: http://arxiv.org/abs/2208.04024.

[6] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. RewriteLM: An instruction-tuned large language model for text rewriting. *Computing Research Repository*, arXiv:2305.15685, 2023. Available from: http://arxiv.org/abs/2305.15685.

[7] Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. *Computing Research Repository*, arXiv:2307.00524, 2023. Available from: http://arxiv.org/abs/2307.00524.

[8] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. Goal-driven explainable clustering via language descriptions. *Computing Research Repository*, arXiv:2305.13749, 2023. Available from: http://arxiv.org/abs/2305.13749.

[9] Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *Computing Research Repository*, arXiv:2302.14233, 2023. Available from: http://arxiv.org/abs/2302.14233.