AI-Curated Democratic Discourse

JSALT workshop 2024 topic – November 2023 [slides, shorter doc, longer doc]

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

We hope to find interface elements that could support public discourse—on news sites like NYT, (micro)blogging sites like X, discussion sites like Reddit, review sites like Yelp, and government agency sites that are required to solicit public comment. They could also be used on social sites like Facebook where most posts and replies are directed to the poster's friends, but some deserve a wider audience because they touch on political or cultural matters of wider interest.

What Will The Interface Do?

We will use generative AI to

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

This design goes beyond the traditional model (Usenet, Reddit, Facebook, X, Nextdoor) where each post is only seen within its original conversation. Although threaded conversations are still central to the experience, their posts are now situated in a broader curated landscape of viewpoints and supporting information.

User interfaces shape user behavior. In this new landscape of argument, 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 also help users as they write their posts, by previewing simulated reactions and offering suggestions before they submit.

Why Is This Important?

In the modern world, everyone can speak publicly. The Internet should enable a broad civic conversation. But scale makes this challenging. Can good ideas rise to the top, or are they confined to narrow circles or drowned out? Especially for matters of consequence, it is important to curate this enormous discourse so that *new ideas* have a chance to be *both heard and appropriately challenged*—and not only by a small set of friends and followers. Curation also reveals the range of views in the community to citizens, policymakers, and journalists.

Without curation, the alternative seems to be a fragmented public discourse, where everyone sees primarily the views of their own "tribe"—including unchallenged misinformation and mischaracterization of other "tribes." A loss of mutual understanding and trust is dangerous for democracy. The public discourse is also mediocre fare, with the most useful and interesting posts being drowned out by the many redundant, unproductive, or polarizing posts.

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 [8, 9], describing differences between different types of text [10], evaluating conversational text on multiple subjective criteria [1, 4], and rewriting text [7].

We anticipate using LLMs for tasks such as these, where P is a post or draft post (note that we consider replies to be posts as well):

- 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 Q that make the same argument as P, so that they should be grouped with P.[1]
- Finding posts Q that consider the same question as P, but provide other 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.

[1] This may involve reading P in context, extracting its claims and arguments in natural language, then producing a vector embedding.

An initial UI sketch, with annotations, is shown in Figure 1. In the left column, the user sees an ordinary threaded social media discussion among individuals. That discussion proceeds vertically as usual—but we use the other two dimensions to augment the display with other material to explore.

Better Reading

We envision a *non-coercive* interface that simply makes additional relevant information available. That is, our threaded reader (Figure 1) will *also* give access to related posts from other threads. The user is free to explore these if they like.

If a post P is on a public matter, we display it atop a *stack* of similar public posts from across the site. While the ordinary replies to P are shown beneath it as usual, other replies to P's stack are shown to its right (when P is selected). These are also grouped into stacks, and are lightly rewritten by AI to make sense as replies to P. One function of this is to highlight supporting arguments and counterarguments.

The user can also explore P's stack by double-clicking on it. We then expand the stack, splitting it into 4 *diverse* substacks that can be recursively explored further. Each substack shows a *high-quality* post on top. Simply clicking on a substack will navigate to show that topmost post in its original context. Also, the user can click on icons to explore posts of different types (agreements, disagreements, information, personal experiences, jokes, predictions, proposals, statements of values, etc.). These features require hierarchical clustering and classification.

Finally, we will allow the user to explore the stack using an LLM, asking for a summary of the stack or asking other questions about it ("Did anyone discuss insurance costs?").

- By featuring diverse posts, we puncture the filter bubble.
- By featuring high-quality posts, we make the site more interesting and valuable for users.
- By giving the user agency in whether and what to explore, and showing multiple options, we defuse the political criticism that we are telling them what to think.

Better Posting

While the user is writing, they can also navigate to other posts as above. They can drag related posts into their own post (like a quote-tweet). This cross-fertilizes discussions among different groups, making it easy for the user to bring in substantive support for their position. It also incentivizes high-quality posts, since authors are told when they've been quoted.

If the user's draft post D appears to be public-facing, we will display it *as if it had already been posted* (though in a different color). The idea is to subtly preview for the user how their post will be received. That is, we will show

- D's entire stack, allowing the user to explore posts similar to D.[2]
- Replies and reactions to D's stack, making the user aware of objections to D's claims or argumentation.

We will also experiment with showing

- Simulated replies to D. This can warn the user about unexpected reactions to the specific phrasing of D.
- An assessment of how inclined the system will be to share D with others, and/or verbal advice on how to improve this metric. (This is an opportunity to encourage posts that are fair, clear, appropriate, and constructive.)

Recommendation Metrics

The user interface tries to improve conversations among humans, in a viewpoint neutral way. But which human-authored posts should we show to a user? We plan to consider these criteria (some depend on user or context):

- quality (factual, well-written, fair-minded, evidence-based, timely)
- enjoyability (interesting, funny, engaging, emotionally satisfying)
- relevance to current context
- coverage (diverse views/groups, novel topics/proposals, key rebuttals)
- representativeness (give readers a sense of which views have more adherents)
- personal connection (see people you know / follow / have engaged with before)
- predicted effect on the user
  - engagement (will they explore, react, reply)
  - effects on subsequent discussion (will it affect their subsequent posts)

[2]This allows a mode where the user starts typing thoughts about a book simply in order to find other conversations about that book. Then the user abandons their draft post and joins an existing conversation instead.

Building the Site

We plan to create an initial site before the start of the workshop, in the form of a modified Mastodon server, which will allow readers to read and post to a subset of globally available Mastodon content (known as the "fediverse"). The Mastodon code is open-source.

If additional content is needed, we will consider simulating additional users [6, 5, 3], prompted by political topics on structured debate sites such as Kialo, DebateWise, and Pol.is.

Team

- Jason Eisner is Professor of Computer Science and a Fellow of the Association for Computational Linguistics. He has worked on many relevant topics, including decision-oriented dialogue, large language models, human behavior modeling (including on social media), and voting system design.
- Andy Perrin is SNF-Agora Professor of Sociology, and the author of books including *American Democracy: From Tocqueville to Town Halls to Twitter* and a forthcoming book on evidence in American public debate. He studies the social and cultural foundations of American democratic life.
- Daniel Khashabi is Assistant Professor of Computer Science. He focuses on building NLP systems that are driven by reasoning and LLMs, including the Perspectrum system [2] for collating perspectives on controversial issues.
- Ziang Xiao is incoming Assistant Professor of Computer Science, who works on HCI. He creates novel conversational interfaces to elicit human preferences, understand human behaviors, survey human opinions, and shepherd political discussion while avoiding backfire effects. His research has informed social media newsfeed design and content moderation mechanisms.

We are working on building a larger and more diverse team. Over the past week, we have been discussing the project with interested relevant researchers at Stanford, CMU, Duke, UCSD, UPenn, Cambridge, and Meta.

Concerns

Q: Isn't AI content bland?
A: We don't plan to post AI-generated content to the site (perhaps with rare exceptions). The goal is to improve conversations among humans.

Q: Won't users feel pushed around?
A: We don't plan to intervene in existing conversations. The current threaded conversation (leftmost column of Figure 1) is still the core of the experience. We simply show other material, which users are fre to explore.

Q: Don't users dislike moderation?
A: Users may dislike being criticized in front of others. However, we are normally giving them advice on their draft before they post it. If social media is gamified, this is telling them privately how to win the game.

Q: Who sets the standards that determine which posts are recommended?
A: There are three separate questions here:
(1) What are the site's values, in human terms?
(2) Can they be reliably scored by humans?
(3) If so, can AI approximate the human scores (e.g., via LLM prompt engineering)?
For the summer workshop, we will attempt a first draft of nonpartisan discourse standards, drawing on past work in political science as well as NLP, responsible AI, and social media, and convening a discussion with a broader group of wise heads. In future, we envision that users could use the platform to debate what standards they want. If necessary, different servers could institute (somewhat) different standards to accommodate users who prefer different experiences.

Q: Aren't there already websites that try to curate political argument?[3]
A: Yes, but they are not social media, where people interact conversationally as individuals. Social media is a significant societal force, so it is important to ensure that it is a force for good, not for ill.

Q: Do users really care about your goals?
A: It is true that users have varying reasons to use social media. (They may want to keep up connections with friends, entertain themselves and others, share reactions, seek confirmation of their worldview, change others' minds, attract attention and followers, or gain social status.) Users are not always actively seeking information and constructive discussion. But we suspect that if we make interesting relevant posts easy to glance at and click on, they may look. Indeed, our AI-curated site should be more pleasurable to read. Scrolling through Internet comments is an addictive waste of time because it provides only intermittent reinforcement. Why wouldn't users prefer to get the highlights?

[3] CrowdLaw highlights a number of efforts such as Pol.is to use technology for deliberative democracy. Kialo and DebateWise are collaborative efforts (like Wikipedia) to produce depersonalized guides to controversial issues.

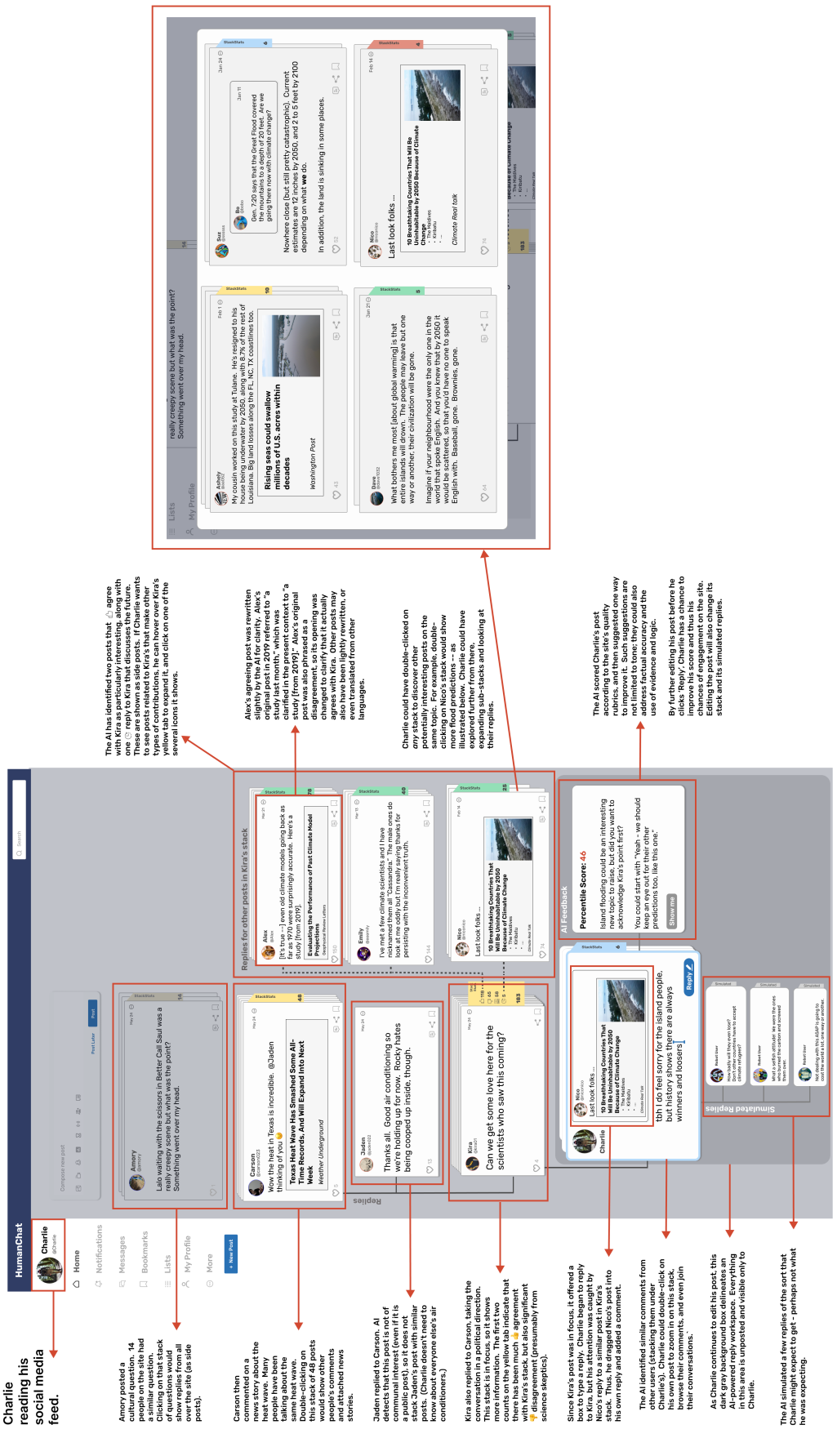

Figure 1: An annotated mockup of our initial UI design.

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] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. 2019. Available from: https://aclanthology.org/N19-1053/.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.