Paper Review: Principles for AI-Assisted Social Influence and Their Application to Social Mediation, Perera et al.

Review by Jason Gillette

Introduction

This paper proposes principles for AI-assisted social influence, emphasizing its application in social mediation as opposed to moderation. Social mediation aims to foster constructive dialogue and resolve disputes, contrasting with moderation's rule-based removal of content. In this context, the authors introduce a social influence system called D-ESC (Dialogue Assistant for Engaging in Social-Cyber-mediation) as a case study to demonstrate these principles within an online setting. For this review, I will provide a brief overview of the major concepts covered followed by several discussion points that may indicate areas for improvement.

What are social influence systems?

Social influence systems refer to frameworks or tools designed to shape individual or community behaviors in alignment with specific goals or ethical considerations. The authors define a successful social influence system as one that must balance ethical transparency, emotional intelligence, and adaptability to dynamic socio-emotional contexts while fostering trust and widespread adoption.

Contributions

The paper introduces two major contributions:

1. **Proposed Principles for AI-Assisted Social Influence**: These six principles address the complexity of ethical AI systems operating in a dynamic social environment such as a Reddit thread containing large scale complex social interactions that evolve over time. The authors emphasize the following principles to guide their research.

    1. Socio-emotional awareness: Understanding and responding to the emotions, beliefs, and values underlying user interactions to align AI interventions with social contexts.
    2. Early intervention: Addressing problematic behaviors proactively before they become habitual or escalate, fostering constructive engagement before problematic dialog becomes entrenched within a community.
    3. Emotional-action linkages: Recognizing how emotions influence behaviors to design interventions that target root emotional triggers rather than symptoms such as toxic language.
    4. Respect for user values: Ensuring interventions operate within individuals' value systems to maximize relevance, acceptance, and effectiveness vice a tone-deaf and thus ignored intervention.
    5. Behavioral modeling: Developing dynamic, personalized models of user behaviors like previously banned posts to predict and influence interactions effectively and adaptively.

6. Transparency: Providing clear, interpretable explanations of AI decisions and actions to build trust and accountability in its use.

2. **D-ESC System**: This system is an application of these principles through a multi-component architecture. It combines natural language processing (NLP), probabilistic reasoning, and sentiment analysis to mediate social interactions effectively. D-ESC claims to detect toxic behaviors, rephrases harmful content, and predicts the long-term impact of interventions on community health.

Methods

The D-ESC system implements NLP modeling techniques which correspond to each principle proposed.

1. **Socio-emotional Awareness**: D-ESC uses stance detection to assess the socio-emotional environment of social media communities. By analyzing expressed and implied emotions, beliefs, and attitudes, the system builds topic-specific lexicons through predicate-argument pairs to refine its moderation and mediation strategies. This automated approach ensures that the interventions align with the nuanced emotions of a given set of online interactions.
2. **Early Intervention**: To predict and shape potentially harmful behaviors, D-ESC incorporates a conversation deviation algorithm. This algorithm, trained using an unsupervised learning approach, identifies early signs of controversy by analyzing chronologically arranged subreddit posts and detecting deviations from expected community norms, enabling timely intervention before conversations deviate from the topic at hand.
3. **Emotional-Action Linkages**: Probabilistic Soft Logic (PSL) models are used to understand how emotions translate into user actions. By defining and learning intervention rules that connect emotional indicators (e.g., anger, sadness) with specific moderator actions, the system creates a mechanism for identifying and addressing covert and overt anti-social behaviors that may lead to problematic interactions.
4. **Respect for User Values**: Leveraging Social Judgment Theory, D-ESC generates rephrasing of inflammatory or toxic posts that align with the acceptable "latitude of acceptance" of targeted communities. These rephrasings preserve the user’s intent while ensuring that the language and tone are aligned with sampled community values.
5. **Behavioral Modeling**: The system uses dynamic modeling techniques, including System Dynamics (SD) Modeling, to simulate interactions between users, post quality, and moderation activity. This enables the creation of a "digital twin" of communities to predict and analyze the outcomes of interventions, tailoring strategies to specific communities and their evolving dynamics.
6. **Transparency**: To ensure trust and adoption, D-ESC provides explainable intervention models that translate PSL outputs into human-readable natural language explanations. Moderators can access these explanations through an interactive dashboard, which also allows them to approve or reject actions, ensuring human-in-the-loop oversight.

Each of these methods are adapted to the online community they serve enabling dynamic response.

Discussion

The discussion highlights several critical challenges and implications:

1. **Unfair Penalties on Dissent**: D-ESC's reliance on moderator-labeled data risks perpetuating biases against dissenting views, which might be flagged as toxic or disruptive for a given online community.

This could suppress constructive disagreement and diversity of thought within communities and be counter-productive toward the author's original intent.

2. **User Adoption of Mediation**: Users and moderators may dismiss AI interventions, particularly if they perceive them as lacking contextual understanding of the discussion or as infringing on free expression. Building trust will require demonstrating consistent fairness and transparency while also showing a convincing level of understanding for the topic at hand. This is analogous to a child being dismissed from an "adult conversation" and represents a significant obstacle for the proposed system to overcome.
3. **Lack of Benchmarking**: While the authors demonstrate rigorous comparisons to previous works for each component of their system, the depth of evaluation only goes as far as simple accuracy metrics of component tasks. It appears no meta-benchmark exists for a system wide comparison nor is there an in-depth discussion on empirical measures of human performance for similar complex mediation. As a result, deployment of the proposed methods may yield anecdotal results and lack the empirical evidence supporting the author's success criteria.

Despite these challenges, D-ESC presents a novel system for moving beyond rule-based moderation to more dynamic mediation which may prove to be a more effective intervention for online communities.

Reference

Perera, I., Memory, A., Kazakova, V. A., et al. (2024). Principles for AI-Assisted Social Influence and Their Application to Social Mediation. *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, 129–140.