9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

The Language of Opinion Change on Social Media under the Lens of Communicative Action

Keywords: Opinion Change, Habermas, Reddit, Communicative Action, NLP

Motivation and setup

Which messages are more effective at inducing a change of opinion in the listener? We approach this question within the frame of Jürgen Habermas' theory of Communicative Action [1], which posits that the *intent* of the message (its pragmatic meaning) is the key. By loading language with intent, the speaker exercises an illocutionary force that can effectively change the hearer's mind based on a shared understanding of reality. This process can be triggered by potentially many different types of illocutionary forces, but especially by virtue of *"shared knowledge, mutual trust, and accord with one another"* [2].

Thanks to recent advances in Natural Language Processing, we can operationalize this theory by extracting latent *social dimensions* that signal intent in natural language. In particular, we use a recently-developed transformer-based classifier that can label conversational text according to the social dimensions it conveys [3]. The tool can identify up to ten fundamental dimensions of the pragmatics of language that have been extensively studied in social science research [4] (e.g., social support, knowledge exchange, expressions of trust).

To identify key ingredients to opinion change, we look at 46k posts and 3.5M comments on Reddit's `r/ChangeMyView`, an on-line forum where people post their opinions and invite others to submit comments that can change those opinions. The poster marks opinion-changing comments with a flag called *delta* (Δ), which we treat as a ground-truth of opinion change.

Results

To find whether expressions of social intent matter in the process of opinion change, we compute the odds ratios of a social dimension d being conveyed by comments with Δ. Comments that express no intent exhibit an odds ratio of 0.23, meaning that they are about 77% less likely to change the mind of the recipient, compared to comments that convey at least one social dimension (Figure 1a). Comments that received a Δ are exceedingly more likely to convey *knowledge* than those with no Δ ($+119\%$). Successful comments are 80% more likely to allude at *similarity* between the stance of the poster and the commenter (e.g., *"I'm glad to know we agree on this"*) or between their experiences (*"My friends used to live in a large city in Asia too"*), and 65% more likely to contain language that discloses *trust* towards entities relevant to their argument (*"I believe what they're saying"*).

Posts by people who end up changing their view are characterized by different social dimensions from those found in view-changing comments (Figure 1b). In particular, those posts are 46% more likely to convey *status*—words of appreciation or gratitude that indicate a respectful approach to the dialogue (*"I have nothing but the utmost respect for service men and women, but ..."*). Conversely, people who introduce their opinion by appealing to power or mentioning power dynamics (*"If people were required to vote, they would take more of an interest in the political situation"*) are the least likely to grant a Δ.

Figure 1c shows a matrix of interaction between the intent of the poster and that of the commenter. Cells represent the variation of the probability of achieving a Δ given a specific combination of intents. For five social dimensions, comments that receive a Δ are more likely

to express an intent that matches that of the original poster. For example, when a post intends to convey a *power* dynamic, the most effective response is to make a similar appeal to power (+22% chance of Δ). This observation is in line with the interpretation of conversations as social exchanges that occur under the assumption that a contribution of a certain type should be matched by a response of a similar type [5]. Some combinations of dimensions that break this symmetry are less likely to reach an agreement. For example, when the poster expresses *power*, comments replying with *status* are 14% less likely to receive a Δ.

To confirm the significance of all these results, we apply logistic regression models by using whether the comment received a Δ as the dependent variable, and all the dimensions and a number of controls for confounders (e.g., activity, political alignment) as independent variables (not shown for brevity). The results are stable, robust to the confounders used, and in line with Habermas' theory.

Discussion

By leveraging recent advances in Natural Language Processing, our work provides an empirical framework for Habermas' theory. Among the various social dimensions we can measure, the ones that are most likely to produce an opinion change are knowledge, similarity, and trust, which resonates with the theory of Communicative Action. However, other dimensions that have not previously been identified as relevant also have a positive effect, such as appeals to power or empathetic expressions of support.

These results, combined, point towards an extension of the original Habermasian theory which includes a more faceted understanding of intent, interpreted as social dimensions of language. The original theory is quite broad in scope, but only provides a few concrete examples. This study contributes to materializing it by providing an empirical descriptive framework for it and finding concrete examples of its effects in the wild.

Our work is a starting point to improve current models of opinion dynamics. The theory of Communicative Action is in stark contrast with how opinion dynamics has been traditionally operationalized, as they describe social interactions as one-dimensional events [6]. These models have been necessarily oversimplified due to the complexity of quantifying social interactions. However, thanks to recent advances in NLP, we are now able to operationalize these concepts and measure from conversations the social intents that are most relevant to the process of opinion diffusion.

References

[1] J. Habermas, *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society* (MIT Press, Cambridge, Mass, 1962).

[2] J. Habermas, *Communication and the Evolution of Society*, vol. 572 (Beacon Press, 1979).

[3] M. Choi, L. M. Aiello, K. Z. Varga, D. Quercia, *Proceedings of the Web Conference 2020* (2020), pp. 1514–1525.

[4] S. Deri, J. Rappaz, L. M. Aiello, D. Quercia, *Proceedings of the ACM on Human-Computer Interaction* **2**, 1 (2018).

[5] P. M. Blau, *Exchange and Power in Social Life* (Transaction Publishers, 1964).

[6] M. Grabisch, A. Rusinowska, *Games* **11**, 65 (2020).

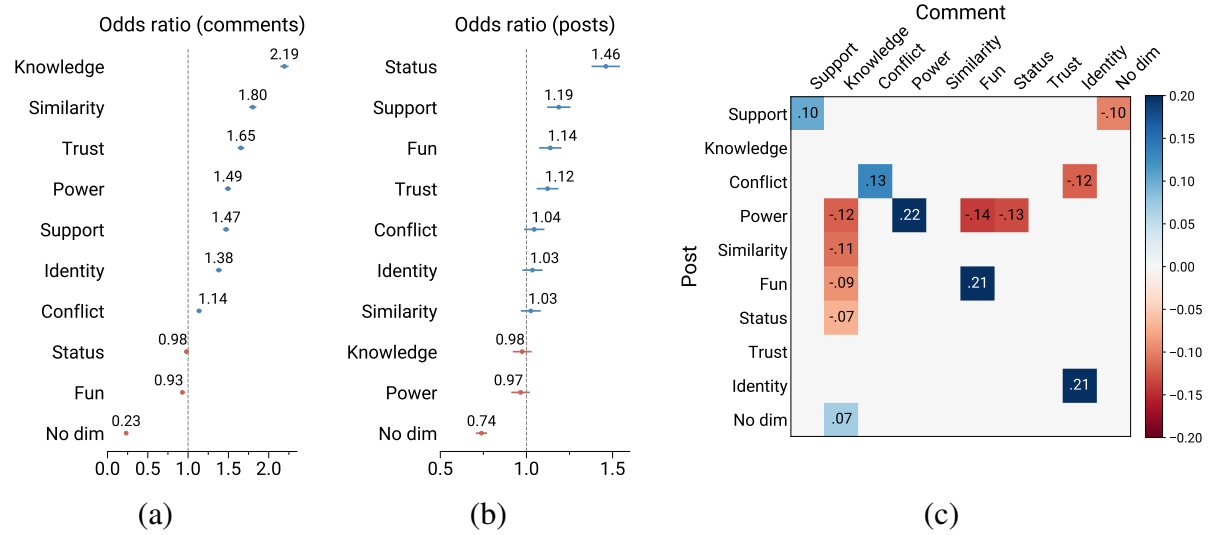

Figure 1: Odd ratios of containing a dimension (a) in comments that were successful in changing the poster's opinion versus those that were not, and (b) in posts expressing opinions that were changed by other community members versus posts that did not experience any opinion change. Error bars represent 95% confidence intervals. On the right (c), we report only the statistically significant odds ratios ($p < 0.01$) for interactions between dimensions in comments and posts. Cells represent the variation of the probability of achieving a Δ given a combination of dimensions.