Computational Models of Argument
H. Prakken et al. (Eds.)

doi:10.3233/FAIA200537

PEOPLES: From Private Responses to Messages to Depolarisation Nudges in Two-Party Adversarial Online Talk

Iwan ITTERMANN [b] and Brian PLÜSS [a]
[a] *Digital Peace Talks gUG (h.b.), Germany*
[b] *Centre for Argument Technology, University of Dundee, UK*

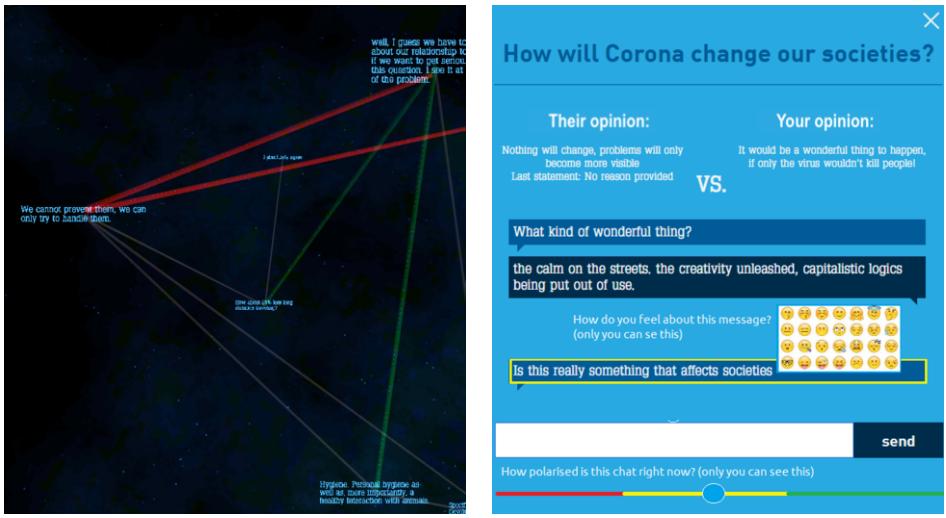
Keywords. Affective Polarisation, Dialogue Modelling, Nudging

The PEOPLES (Private Expression of Polarisation Leveraged to Expand Sociability) Project envisages a fine grained, language-independent measure of affective polarisation between participants in two-party chats over controversial topics. The ultimate goal of the project is to channel the analytical power of the measure to enable automated real-time interventions, nudging participants towards healthier conversational behaviours.

We hypothesise that this measure can be derived solely from the unique profiles of each conversational participant's private reactions (akin to emoji responses on mainstream social media) to the messages they receive in two-party chats. Aided by the language-independence of the approach, we intend to base and evaluate the measure on empirical evidence, by studying polarised users from several cultural contexts, both Western and non-Western.

So far, much emphasis has been on text classification to detect hate speech [1,2], profanity [3] and incivility [4], or on sentiment analysis and psychometric measuring to identify influential factors on political polarisation in deliberative spaces and networks [5,6,7]. Both approaches have limitations when it comes to developing helpful automated interventions at scale. The former assumes uniform reactions across all participants and is thereby prone to have discriminatory effects on minorities, while depending on substantial, costly training datasets. The latter is descriptive, assuming polarisation to be the effect of actions (e.g. news consumption, media use) or connectivity (network popularity, group contact), thus offering little insight for effective automated interventions.

To the best of our knowledge, researchers have not previously employed opinion polarisation analysis based on two-party private communication such as chats online. One of the main reasons for this is the scarcity of natural data publicly available, due to privacy constraints. DPT (`demo.dpt.world`) offers an uncommon opportunity to access such data: it publishes and structures two-party discussions between opinion postings in a signed graph (see Figure 1a). Conceptually, it is comparable to ChangeAView (`changeaview.com`), with the difference that users are required to post their opinions regarding a given topic before they can take part in one-to-one discussions. Chat messages are published after three days. Users can continuously rate the chat's degree of polarisation. The averaged ratings of both users determine the weight of the edge connecting both postings in the graph. In a new feature, participants of a chat will be able



(a) Graph of opinions (nodes) and chats between posters (edges).

(b) DPT chat interface with emoji reaction to messages and polarisation rating.

Figure 1. The PEOPLES-DPT system

to click on icons (comparable to emoji reactions in Messenger, only they are not visible to the other party during the conversation) to privately record their reaction to a specific message (see Figure 1b).

The exploration of a sender-receiver aware polarisation measure, as well as receiver aware nudge-style interventions, is aimed at advancing the understanding of the role of messengers in affective opinion polarisation, and at laying ground for depolarisation technologies to gain momentum.

References

[1] S. Akhtar, V. Basile and V. Patti, A New Measure of Polarization in the Annotation of Hate Speech, in: *AI*IA 2019 – Advances in Artificial Intelligence*, M. Alviano, G. Greco and F. Scarcello, eds, Springer International Publishing, Cham, 2019, pp. 588–603. ISBN ISBN 978-3-030-35166-3.

[2] P. Fortuna and S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* **51**(4) (2018), 1–30.

[3] P.L. Teh, C.-B. Cheng and W.M. Chee, Identifying and Categorising Profane Words in Hate Speech, in: *Proceedings of the 2nd International Conference on Compute and Data Analysis*, ICCDA 2018, Association for Computing Machinery, New York, NY, USA, 2018, pp. 65–69. https://doi.org/10.1145/3193077.3193078.

[4] T. Hopp, C.J. Vargo, L. Dixon and N. Thain, Correlating self-report and trace data measures of incivility: A proof of concept, *Social Science Computer Review* (2018), 0894439318814241.

[5] J.N. Druckman and M.S. Levendusky, What do we measure when we measure affective polarization?, *Public Opinion Quarterly* **83**(1) (2019), 114–122.

[6] J. Yang, H. Rojas, M. Wojcieszak, T. Aalberg, S. Coen, J. Curran, K. Hayashi, S. Iyengar, P.K. Jones, G. Mazzoleni et al., Why are "others" so polarized? Perceived political polarization and media use in 10 countries, *Journal of Computer-Mediated Communication* **21**(5) (2016), 349–367.

[7] C.A. Bail, L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.F. Hunzaker, J. Lee, M. Mann, F. Merhout and A. Volfovsky, Exposure to opposing views on social media can increase political polarization, *Proceedings of the National Academy of Sciences* **115**(37) (2018), 9216–9221.