Rehumanisation in symbolic space of online social media: NLP approach to detection of rehumanising acts

Keywords: rehumanisation, blatant dehumanisation, Natural Language Processing, social media, online abuse

Extended Abstract

There is a distinct public awareness and recognition of the problem of social polarisation. That social polarisation is both reflected in and triggered by between-group online abuse. This concern is also reflected in a large and growing body of scientific literature on online hate speech and dehumanising language. However, researchers rarely focus on how groups with a history of hostility and confrontation come back to respectful dialogue and what structural, social and technological factors facilitate between-group reconciliation.

The internet, and social media in particular, are not only a place of expression of attitudes between groups, but also a means for shaping them. Thus, understanding the respective mechanisms and knowledge of the ways to overcome online polarisation and radicalisation are relevant for successful social governance.

Motivated by these considerations, this study is focused on the topic of rehumanisation in online interactions. The study analyses the language used when describing the outgroup, as this can shift from hostile to accepting. The most striking manifestation of inter-group hostility is the relationship between nations at war. Therefore, the case study for the proposed research is the relationship between Ukrainians, Russians, and Belarusians after the Russian invasion of Ukraine in February, 2022 as reflected in online social media interactions amongst the national groups of these three countries.

Dehumanisation is known to hinder empathy (Bastian et al., 2011), undermine prosocial behaviour (Wohl et al., 2012), and contribute to between-group aggression (Viki et al., 2013). Consequences of dehumanisation are particularly dangerous in the context of international military conflicts, as it can contribute to support for war (Jackson & Gaertner, 2010) and war-related violence (Viki et al., 2013), and support for political violence and extreme policies against the outgroup in countries with a history of military confrontation (Leidner et al., 2013; Maoz & McCauley, 2008). Detecting sites of out-group dehumanisation and finding approaches to overcome it is no easy task, but instruments of Natural Language Processing (NLP) bring new opportunities to address this problem (Mendelsohn et al., 2020).

The research proposed for the presentation consists of two studies. Study 1 focuses on theoretical work and generates the framework for computational analysis of rehumanisation in online interactions. The study will be finished by the time of the conference. Study 2 represents a data-driven work and aims to operationalise rehumanising act in online communications using NLP tools. The author will share formulated hypotheses and preliminary results of pilot analysis of this second study.

Study 1 proposes a framework for analysis of rehumanising activity in the context online communications and answers two research questions: “How rehumanisation reveals itself in online communication?” and “What are its measurable markers?”. First, building on literature on rehumanisation and related topics, the author develops a theoretically motivated definition

of a rehumanising act in symbolic space of online social media. Second, the author collects a database of several dozens of rehumanising acts from two sources: online media outlets and Youtube blogs discussing social, political and military events in Ukraine, Russia and Belarus. Third, utilising digital ethnography approach, they conduct a qualitative analysis of the collected rehumanising acts. This will inform hypotheses for Study 2.

Study 2 utilises NLP tools adopted for Russian, Ukrainian and Belarusian languages to answer the research question "How can rehumanising act be automatically identified from language use of online communications?". Rehumanising act reveals itself through specific components such as empathy to outgroup, highlighting their moral agency, telling about between-group help etc. Many of these components may be identified in text. Explorative work to be completed in Paper 1 will inform on particular instruments to be tested as rehumanisation indicators in Paper 2. For example, one may try using extension of Connotation Frames for agency (Sap et al., 2017) and Pattern-based measure of empathy (Montiel-Vázquez et al., 2022) to measure respective components of a rehumanising act. The dataset for the Study 2 will be created by collecting textual data from the selected list of relevant online media outlets that discuss social, political and military events in the region. Finally, utilising deductive approach and using NLP tools the study will develop a predictive classification model to automatically identify rehumanising acts in text.

References

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status: Dehumanization and moral status. *British Journal of Social Psychology*, *50*(3), 469–483. https://doi.org/10.1348/014466610X521383

Jackson, L. E., & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive Behavior*, *36*(4), 238–250. https://doi.org/10.1002/ab.20344

Leidner, B., Castano, E., & Ginges, J. (2013). Dehumanization, Retributive and Restorative Justice, and Aggressive Versus Diplomatic Intergroup Conflict Resolution Strategies. *Personality and Social Psychology Bulletin*, *39*(2), 181–192. https://doi.org/10.1177/0146167212472208

Maoz, I., & McCauley, C. (2008). Threat, Dehumanization, and Support for Retaliatory Aggressive Policies in Asymmetric Conflict. *Journal of Conflict Resolution*, *52*(1), 93–116. https://doi.org/10.1177/0022002707308597

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, *3*(August), 1–24. https://doi.org/10.3389/frai.2020.00055

Montiel-Vázquez, E. C., Ramírez Uresti, J. A., & Loyola-González, O. (2022). An Explainable Artificial Intelligence Approach for Detecting Empathy in Textual Communication. *Applied Sciences*, *12*(19), 9407. https://doi.org/10.3390/app12199407

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., & Choi, Y. (2017). Connotation Frames of Power and Agency in Modern Films. *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, 2329–2334. https://doi.org/10.18653/v1/D17-1247

Viki, G. T., Osgood, D., & Phillips, S. (2013). Dehumanization and self-reported proclivity to torture prisoners of war. *Journal of Experimental Social Psychology*, *49*(3), 325–328. https://doi.org/10.1016/j.jesp.2012.11.006

Wohl, M. J. A., Hornsey, M. J., & Bennett, S. H. (2012). Why group apologies succeed and fail: Intergroup forgiveness and the role of primary and secondary emotions. *Journal of Personality and Social Psychology*, *102*(2), 306–322. https://doi.org/10.1037/a0024838