Sarah E. Cornwell
University of Western Ontario, London, Ontario, Canada

Nicole S. Delellis
University of Western Ontario, London, Ontario, Canada

Dominique Kelly
University of Western Ontario, London, Ontario, Canada

Yifan Liu
University of Western Ontario, London, Ontario, Canada

Alex Mayhew
University of Western Ontario, London, Ontario, Canada

Yimin Chen
Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia

Victoria L. Rubin
University of Western Ontario, London, Ontario, Canada

THEORIZING IMPROVED NLP FEATURES FOR PROMOTING BEHAVIOR THAT SUPPORTS CMC USERS' SUBJECTIVE WELL-BEING

Abstract

This work-in-progress identifies gaps in the current Natural Language Processing (NLP) approaches for pro-social communication detection by organizing the state-of-the-art NLP feature detection approaches according to models of Subjective Well-Being (SWB) from Positive Psychology. We need to better understand the current state of the field and what features of prosocial computer-mediated communication (CMC) we have yet to address.

Literature Review

State of NLP Detectors and Promotion of Positive Behaviour

Hate speech, verbal victimization, and other negative interactions in computer-mediated environments have been targeted by NLP detection systems for over a decade (e.g., Dinakar et al., 2012). Newer approaches include the identification of positive social interactions in order to promote their visibility or encourage positivity online more proactively (Bao et al., 2021). This sub-field is nascent and growing, with most projects addressing the detection of positive social interactions appearing in the past five years (Sametoğlu et al., 2022). The more established field of sentiment and emotional analysis (e.g., Liu, 2012; Pang & Lee, 2008) has contributed to the development of detectors for prosocial behaviors that include positive emotions, such as laughter or compliments (Bao et al., 2021). As of yet, many of the existing projects have not engaged deeply with psychological theory that describes behaviors which are known to improve individuals' subjective well-being (SWB; roughly akin to life satisfaction). Instead, the current state-of-the-art in NLP has been limited to elements which are easier to compute with current technology, such as the expression of positive sentiment (e.g., detecting compliments or expressions of gratitude) or sustained conversational partnership (e.g., Bao et al.'s (2021) "sustained conversation" feature; Hughes et al.'s (2024) 'responsivity' feature). However, communication with positive individual outcomes (i.e., communication that contributed to higher SWB) is much more diverse than these early efforts have attempted to identify through automation. Automation is likely to be successful at detecting a more holistic range of these behaviors: BERT models applied to prompt responses correlate to SWB self-rating scales with high accuracy (Kjell, 2022); these successes are likely to transfer to more naturalistic language.

Frameworks and Models of Improving Subjective Wellbeing

"Prosocial behavior represents a broad category of acts that are defined by some significant segment of society and/or one's social group as generally beneficial to other people" (Penner et al., 2005, p.366). For example, *gratitude* is one of the primary psychological mechanisms thought to underlie reciprocal altruism (Emmons & McCullough, 2003). Lambert et al. found experimentally that "increasing the regularity and frequency of expressing gratitude enhanced participants' perception of the communal strength of their relationship with their friend" (2010, p.578). There is a clear connection between benefiting others and doing good for ourselves. 'Happiness', referred to as SWB in psychological discourse, is a multi-faceted construct. Diener's (1984) tripartite model of SWB acknowledges that the construct is subjective, includes positive measures, and involves "a global assessment of all aspects of a person's life" (p. 544). The umbrella term of SWB is thought to be "comprised of life satisfaction, which is a broad cognitive appraisal regarding one's life, and affective feelings, which include abundant positive feelings and minimal negative feelings" (Heintzelman & Tay, 2017, p. 10). Since Diener's (1984) widely used and accepted tripartite model, positive psychologists have developed more complex constructs of well-being.

Seligman's (2011) work proposes the PERMA model as encapsulating five elements of well-being that enable flourishing – Positive Emotion, Engagement, Relationships, Meaning and Accomplishment. Each of these five elements meets the three necessary criteria to be distinct from the others: "1) it contributes to well-being; 2) many people pursue it for its own sake, not merely to get any of the other elements; 3) it is defined and measures independently of the other elements (exclusively)" (Seligman, 2011, p. 16). Positive emotion relates to the hedonic conceptualization of happiness: increased feelings of positive emotion. Wagner et al. (2019) effectively summarizes the remaining four elements of PERMA as: "engagement (i.e., being often completely focused and losing the track of time, as in flow experiences; cf. Csikszentmihalyi, 1990), positive relationships (i.e., having close interpersonal relationships), meaning (i.e., having a sense of purpose in life), and accomplishment (i.e., having ambitions, goals, and experiencing mastery)" (p. 309). The five elements of the PERMA model attempt to capture the breadth of factors that can influence well-being. Seligman et al. (2009) proposed the idea of teaching well-being in schooling system to facilitate the students' academic performance and life satisfaction, it has been operationalized through the PERMA(H) and VIA framework in positive psychology. The PERMA(H) model (Norrish et al., 2013) adds positive health to Seligman's model. The PERMA(H) model has been applied in school-based settings as multiple interventions and tests of their effectiveness on students' well-being (Norrish, 2015; Tansey et al., 2018). Kern et al.'s (2015) cross-sectional study suggests that each domain would relate differentially to a range of well-being outcomes.

Discussion

Initial work has started to detect the full spectrum of elements which contribute to SWB (e.g., Bao et al., 2021), but further nuance and incorporation of more advanced NLP methods promises further improvements. For example, Kiesling et al.'s (2018) work on stancetaking could be adapted to identify co-operative efforts, expressions of gratitude more complex than simple "thanks", and disagreements that eventually lead to agreement. Pro-social behaviors that reflect the PERMA elements of meaning, positive relationships, and accomplishment have the potential to promote more pro-social cultures in online spaces.

Improving the detection of linguistic behaviors which promote SWB has several potentially useful applications. Current popular moderation strategies in Information Communication Technologies (ICTs), especially social media, could benefit from not only using a subtractive moderation approach (reducing negative interactions) but from a more nuanced additive moderation approach (being able to detect and promote a wider range of complex positive interactions). Subtractive moderation has problems besides censorship. For example, errors by algorithms which remove hate speech can exacerbate harms against marginalized groups: false positives can shut down marginalized groups' discussions about or descriptions of hate speech, while false negatives allow hate speech to remain on a platform (Davidson et al., 2017). Promoting positive interactions may be less likely to have this kind of harmful knock-on effect, though there is the risk that especially salient incentives (e.g., monetary rewards) for completing positive behaviors can decrease users' internal altruistic motivations (Qiao, 2017). Even when strictly applied, subtractive moderation does not always remove the potential to do harm. Strict design features, such as severely limiting linguistic freedom by maintaining a short list of allowable types of communicative user interaction, (e.g., as in mobile game *Hearthstone*) does not necessarily eliminate negative interactions (e.g., *Hearthstone*'s "sorry" emote became a *de facto* negative linguistic message due to being spammed sarcastically[1]). Experimental promotion of 'prosocial' news comments has shown promise in improving readers' perception of comment sections as informative and interesting (Saltz et al., 2024); this direction deserves further work.

[1] See: https://www.eurogamer.net/blizzard-has-nerfed-hearthstones-sorry-emote Accessed January 20, 2025

Further improvements could be developed by investigating the effects of non-linguistic features including social networks and affordances allowed by design features. Research has begun to link design features of social media platforms to affordances that they enable (e.g., Treem & Leonardi, 2012; Kelly et al., 2022). Affordances are the possible actions that emerge from the relation between the properties of an object and an interacting agent's capabilities (Norman, 2013). Future work could determine whether pro-social behavior in online communities, detected through NLP methods, is associated with the existence of certain affordances. If so, the features that enable those affordances could be recommended for use in social media platforms, contributing to a line of research that examines how technologies can be designed to promote users' SWB (Desmet & Pohlmeyer, 2013; Riva et al., 2012). Non-linguistic SWB promoting affordances have been identified, including those that aim to keep users on platforms longer (Monge Roffarello et al., 2024).

Modern ICTs have been found to have negative mental health effects (Sadagheyani & Tatari, 2020). A better understanding of the effects of linguistic behaviors on participants' SWB would allow for the promotion of outcomes with better pro-social outcomes and higher SWB of participants. Positive communication behaviors such as greeting and complimenting can promote feelings of joy, happiness, and gratitude (Mirivel, 2019, p. 57). Expressions of gratitude have been found to increase pro-social behavior (Grant & Gino, 2010) and there is evidence that such expressions may increase the subjective well-being of both the person thanking and being thanked (Yoshimura & Berzins, 2017). Gratitude detection has begun to be automated (Bao et al., 2021), but the transferability of these findings to user outcomes has not been thoroughly investigated. With careful thought into the ethics of promoting specific behaviors and methodical testing with an ethically diverse group of participants, it may be possible to build ICTs which help promote positive mental health outcomes and greater SWB in their userbases (Mayhew et al., 2022).

Conclusion

This work contributes towards identifying a promising synergy in research within LIS&T: detecting and promoting prosocial messages aligns with librarianship's pro-social goals while building on the existing body of sentiment analysis research. The detection of positive emotion and life satisfaction is now well advanced in NLP (e.g., Kjell et al., 2022), yet its prosocial application to online communities is not well attested. More theoretical and empirical work is needed in this inter-disciplinary direction to more precisely and holistically detect the full range of SWB elements in online communities. The promotion of PERMA elements like accomplishment, meaning, and positive relationships requires more robust detection abilities.

Acknowledgements

This research has been funded by the Government of Canada Social Sciences and Humanities Research Council (SSHRC) Insight Grant awarded to Dr. Victoria L. Rubin's Language and Information Technology Research Lab (LiT.RL) for the 2024-2028 project entitled "*Well-being*

in Social Media: Applying Natural Language Processing Analyses to Discover Thriving Online Communities, Promote Healthier Social Media Design and Authentic Happiness."

References

Bao, J., Wu, J., Zhang, Y., Chandrasekharan, E., & D. Jurgens. (2021). Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery*, New York, NY, USA, 1134–1145. https://doi.org/10.1145/3442381.3450122

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512-515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Desmet, P. M. A., & Pohlmeyer, A. E. (2013). Positive design: An introduction to design for subjective well-being. *International Journal of Design*, *7*(3), 5-19.

Diener, E. (1984). Subjective Well-Being. *Psychological Bulletin*, *95*(3), 11–58. https://doi.org/10.1007/978-90-481-2350-6_2

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1-30.

Emmons, R.A., & McCullough, M.E. (2003). Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life. *Journal of Personality and Social Psychology*, *84*, 377–389.

Grant, A. M., & Gino, F. (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of personality and social psychology*, *98(*6), 946. https://doi.org/10.1037/a0017935

Heintzelman, S., & Tay, L. (2017). Subjective well-being: Payoffs of being happy and ways to promote happiness. *Positive Psychology: Established and Emerging Issues* (pp. 9–28). https://doi.org/10.4324/9781315106304

Hughes, M., Roy, B. C., & Roy, D. (2024). In Pursuit of Constructive Communication: Designing Tools to Support Development of Constructive Communication Metrics. *Designing Interactive Systems Conference*, 121–124. https://doi.org/10.1145/3656156.3663720

Kelly, D., Liu, Y., Mayhew, A., Chen, Y., Cornwel, S.E., Delellis, N.S. and Rubin, V.L. (2022), Supporting Prosocial Behaviour in Online Communities through Social Media Affordances. *Proceedings of the Association for Information Science and Technology*, *59*: 723-725. https://doi.org/10.1002/pra2.703

Kern, M. L., Waters, L. E., Adler, A., & White, M. A. (2015). A multidimensional approach to measuring well-being in students: Application of the PERMA framework. *The Journal of Positive Psychology*, *10*(3), 262-271. doi: 10.1080/17439760.2014.936962

Kiesling, S.F., Pavalanathan, U., Fitzpatrick, J., Han, X., & J. Eisenstein. (2018). Interactional Stancetaking in Online Forums. *Computational Linguistics*, *44* (4): 683–718. doi: https://doi.org/10.1162/coli_a_00334

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, *12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers. Retrieved from http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.html

Mayhew, A., Chen, Y., Cornwell, S.E., Delellis, N.S., Kelly, D., Liu, Y. and Rubin, V.L. (2022), Envisioning Ethical Mass Influence Systems. Proceedings of the Association for Information Science and Technology, 59: 756-758. https://doi.org/10.1002/pra2.716

Mirivel, J. C. (2019). Communication Behaviors That Make a Difference on Well-Being and Happiness. *The Routledge handbook of positive communication: Contributions of an emerging community of research on communication for happiness and social change*. https://doi.org/10.4324/9781315207759

Monge Roffarello, A., De Russis, L. & Pellegrino, M. (2024). Digital Wellbeing Lens: Design Interfaces That Respect User Attention. In Proceedings of the 2024 International Conference on Advanced Visual Interfaces (AVI '24). Association for Computing Machinery, New York, NY, USA, Article 51, 1–5. https://doi.org/10.1145/3656650.3656674

Norman, D. (2013). *The design of everyday things*. Basic Books.

Norrish, J. M., Williams, P., O'Connor, M., & Robinson, J. (2013). An applied framework for positive education. *International Journal of Wellbeing*, *3*(2). 147-161. doi:10.5502/ijw.v3i2.2

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000001

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. Annu. Rev. Psychol., 56, 365-392. https://doi.org/10.1146/annurev.psych.56.091103.070141

Qiao, D., Lee, S.-Y., Whinston, A., & Wei, Q. (2017). *Incentive Provision and Pro-Social Behaviors*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2017.675

Riva, G., Baños, R. M., Botella, C., Wiederhold, B. K., & Gaggioli, A. (2012). Positive technology: Using interactive technologies to promote positive functioning. *Cyberpsychology, Behavior, and Social Networking*, *15*(2), 69-77. https://doi.org/10.1089/cyber.2011.0139

Sadagheyani, H. E., & Tatari, F. (2020). Investigating the role of social media on mental health. *Mental health and social inclusion*. https://doi.org/10.1108/MHSI-06-2020-0039

Saltz, E., Jalan, Z., & Acosta, T. (2024). *Re-Ranking News Comments by Constructiveness and Curiosity Significantly Increases Perceived Respect, Trustworthiness, and Interest* (arXiv:2404.05429). arXiv. https://doi.org/10.48550/arXiv.2404.05429

Sametoğlu, S., Pelt, D., Eichstaedt, J C., Ungar, L. H., & Bartels, M. (2022). *The Value of Social Media Language for the Assessment of Wellbeing: A Systematic Review and Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qnx2v

Seligman, M. (2011). Flourish: a visionary new understanding of happiness and well-being. Atria paperback.

Seligman, M. E., Ernst, R. M., Gillham, J., Reivich, K., & Linkins, M. (2009). Positive education: Positive psychology and classroom interventions. *Oxford review of education*, *35*(3), 293-311. doi: 10.1080/03054980902934563

Tansey, T. N., Smedema, S., Umucu, E., Iwanaga, K., Wu, J.-R., Cardoso, E. da S., & Strauser, D. (2018). Assessing College Life Adjustment of Students With Disabilities: Application of the PERMA Framework. *Rehabilitation Counseling Bulletin*, *61*(3), 131–142. https://doi.org/10.1177/0034355217702136

Treem, J. W., & Leonardi, P. M. (2012). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Communication Yearbook*, 36. http://dx.doi.org/10.2139/ssrn.2129853

Wagner, Gander, F., Proyer, R. T., & Ruch, W. (2019). Character Strengths and PERMA: Investigating the Relationships of Character Strengths with a Multidimensional Framework of Well-Being. *Applied Research in Quality of Life*, *15*(2), 307–328. https://doi.org/10.1007/s11482-018-9695-z

Yoshimura, S. M., & Berzins, K. (2017). Grateful experiences and expressions: The role of gratitude expressions in the link between gratitude experiences and well-being. *Review of Communication, 17*(2), 106-118. https://doi.org/10.1080/15358593.2017.1293836