A Generative Multi-Agent Framework for Modeling Depressive Language Entrainment

Chris Hoynck van Papendrecht[0009−0006−5190−9304], Johan Bollen[0000−0001−7031−9293], and Debraj Roy[0000−0003−1963−0056]

Computational Science Lab, University of Amsterdam,
Science Park 904, 1098 XH Amsterdam, The Netherlands.
chris.hoynck.van.papendrecht@student.uva.nl,
{j.l.t.m.bollen,d.roy}@uva.nl

Abstract. When people interact, they adjust their language to each other. This process of lexical entrainment fosters social agreeableness, but may also create jargon, hypes, memes, and even political polarization when communities converge on a shared vernacular. Here we study what happens when people with depression interact with each other in online social networks with varying degrees of lexical entrainment. We connect generative AI agents in a social network, each endowed with a personal mental health profile. The agents exchange Tweet-like messages that are shaped by their individual score on a PHQ-9 depression questionnaire, while the exchanges induce lexical entrainment. We find that cognitive distortions, a style of thinking associated with and possibly causative of internalizing disorders, can rapidly diffuse in social networks through the process of lexical entrainment, creating a depressogenic psycho-social environment that may lead to worse mental health outcomes throughout the community. Our results may inform targeted approaches to remove risk factors associated with social media use and mitigate the effects of lexical entrainment in communities with mental health challenges.

Keywords: Lexical Entrainment · Depression Contagion · Generative Agent-Based Modeling

1 Introduction

The use of language on social media has transformed how emotions, beliefs, and cognitive patterns spread among individuals. Human communication is inherently adaptive: people subconsciously adjust their lexical and syntactic choices during interaction, a phenomenon described by accommodation and convergence–divergence theories [19]. Lexical convergence leads to lexical entrainment, through which social groups can develop a shared vernacular and vocabulary that fosters social bonding and perceived trust, often strengthening in-group identity [6]. However, the same mechanisms that forge cohesion can also amplify maladaptive patterns. When communities develop shared vernaculars around pessimistic or distorted expressions, linguistic convergence may inadvertently

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

reinforce collective negativity—mirroring processes observed in echo chambers, meme diffusion, and political polarization.

Depression, in particular, is closely tied to linguistic expression [3]. Cognitive distortions—habitual patterns such as catastrophizing, overgeneralization, or all-or-nothing thinking—shape both how individuals talk and how they interpret social feedback. Cognitive Behavioral Therapy (CBT) interventions explicitly target these distorted linguistic markers to modify underlying thought patterns. Empirical studies have demonstrated that increased use of depressive language predicts worsening mental health outcomes and higher PHQ-9 scores over time [9,3]. Crucially, exposure effects vary: not everyone who encounters distorted language adopts it to the same degree. Personality traits, baseline mood, and network centrality jointly determine susceptibility to emotional contagion [10].

The rise of social media introduces a new scale and structure of psychological exposure. Unlike broadcast media, social platforms operate as complex feedback systems in which individuals are both consumers and producers of emotionally valenced language [11]. Peer-generated messages—amplified by likes, replies, and retweets—circulate through densely connected, overlapping communities. Yet despite abundant data on depressive content online, our understanding of how cognitive distortions propagate through these networks remains limited. Existing research largely focuses on individual outcomes or text classification, overlooking the emergent collective patterns that may sustain or intensify depressive discourse [14,20].

Generative Agent-Based Modeling (G-ABM) offers an innovative approach to bridge this gap [15]. Generative agents have proven to be effective proxies for both human behavior and diverse demographic groups, with silicon societies further demonstrating the collective emergent conventions observed in real-world populations [15,1,2]. Built upon principles of lexical accommodation and network diffusion, this method enables simulation of linguistic adaptation at both local and global scales. Empirical studies suggest that agents powered by large language models (LLMs) reproduce realistic patterns of language style matching, providing a uniquely powerful tool for modeling social cognition [8]. Within this framework, linguistic entrainment can be quantified using embedding similarity, n-gram divergence, or semantic drift over time[4,16]. At the network level, diffusion dynamics depend on topology—where networks with highly connected nodes may induce nonlinear, threshold-like adoption of depressive language.

Despite progress in computational psycholinguistics and social network analysis, key knowledge gaps remain. How do micro-level affective expressions aggregate into macro-level linguistic climates of depression? Which network configurations amplify or attenuate the spread of cognitive distortions, and are there identifiable tipping points where the process becomes self-sustaining? Understanding these threshold dynamics is critical for identifying intervention leverage points—network positions or interaction rules that can halt maladaptive diffusion.

To address these questions, we develop a Generative Agent-Based Model in which autonomous agents interact through a simulated social media platform.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

Each agent perceives, rephrases, and posts content using LLM-based generation conditioned on internal affective states. We leverage their diagnostic capabilities for inference of mental state [13]. The network topology follows the Social Distance Attachment Model [18], capturing homophily-driven connections that mirror real-world social structures. By tracing well-being assortativity and sentiment drift across interactions, we aim to uncover the conditions under which depressive language becomes contagious and, ultimately, how linguistic interventions might disrupt these feedback loops.

2 Method

2.1 Models and Data

This project utilizes meta-llama/Llama-3.1-8B-Instruct to drive generative dialogue [7]. To quantify language entrainment we use MentalBERT to capture the nuances of mental health discourse [12]. Our datasets rely on PersonaHub to generate synthetic "silicon personas" for scalable interactions [5]. The psychological profiles and ages of these personas are sampled using a Dutch database categorized by the PHQ-9 depression screening tool [17].

2.2 Simulation

To study how language drifts within a social network, we simulate a population of artificial agents that mirror the behavior of social media users. Each agent is defined by a specific persona and a dynamic well-being state, represented by a PHQ-9 score.

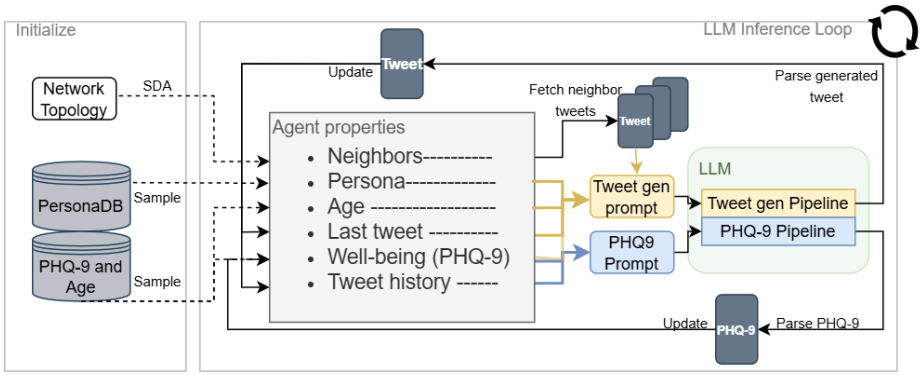

Fig. 1. Agent-based simulation pipeline for tweet generation and PHQ-9 assessment.

The Social Distance Attachment (SDA) topology embeds agents in a d-dimensional social space constructed from latent geometric coordinates and normalized empirical attributes. In the SDA model, the connection probability between agents

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

i and j is defined as [18]:

$$p_{ij} = \left[1 + \left(\frac{d_{ij}}{b}\right)^{\alpha}\right]^{-1} \tag{1}$$

where d_{ij} is the Euclidean distance, α regulates homophilic strength, and b is a scale parameter calibrated via bisection to satisfy a target expected degree k.

The simulation follows a discrete-time loop, depicted in Figure 1, which synchronizes agent state updates. The process transitions through the following stages:

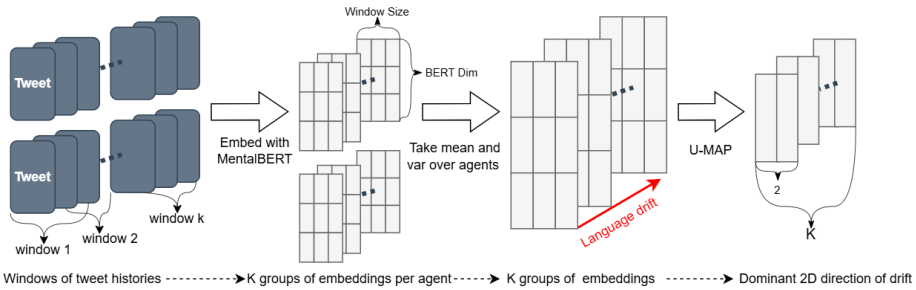

Fig. 2. Process of calculating language drift using Mental-BERT embeddings and U-MAP dimensionality reduction.

- **Mental Health Assessment (Blue Pathway):** Periodically (every x rounds), agents undergo a PHQ-9 evaluation. The system assembles a context comprising the agent's recent tweet history and previous well-being scores. This context is processed by an LLM, which infers the agent's depressive symptoms and mood based on the content and linguistic tone of their activity.
- **Social Interaction and Content Creation (Yellow Pathway):** To generate a new tweet, the agent integrates their persona, age, and current well-being score with a social context consisting of their own last post and the neighbors' tweets from the previous round. Using this context, the agent evaluates their activation to determine if they are motivated to post in the current step. If activated, the LLM generates a tweet tailored to the agent's profile and its surrounding social environment.
- **Synchronous State Update:** To maintain temporal consistency and eliminate informational advantages, the network state updates globally at the end of each step. This ensures that all neighbor interactions and post-visibilities are synchronized for the subsequent round.

Here, we simulate with a population of 200 agents for 600 iterations. We set $\alpha = 1$, and for the distance matrix utilize a 2-dimensional Gaussian social space

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

based on initial PHQ-9 and age as empirical normalized attributes to define the Euclidean distances d. We experiment with degree $k = 6$ and $k = 9$.

2.3 Metrics

Macro-level language dynamics are captured by tracking semantic drift over time using a structured pipeline (illustrated in Figure 2). We apply MentalBERT embeddings to rolling windows of agent tweets. We project these high-dimensional embedding shifts into a localized 2D latent space using UMAP. This dimensionality reduction visually maps the dominant trajectories of language adoption.

Dynamic, degree-weighted (DW) PHQ-9 scores serve as our complementary psychological metric. We continuously estimate individual depressive severity by averaging PHQ-9 scores derived from each agent's tweet history. To identify clustering and depressive echo chambers, we compute well-being assortativity—using the Pearson correlation coefficient—between individual and DW PHQ-9 scores. Finally, we calculate the cross-agent variance of the DW PHQ-9 to detect the formation of multiple distinct clusters, aggregating these metrics across time steps to visualize general temporal trends.

3 Results and Discussion

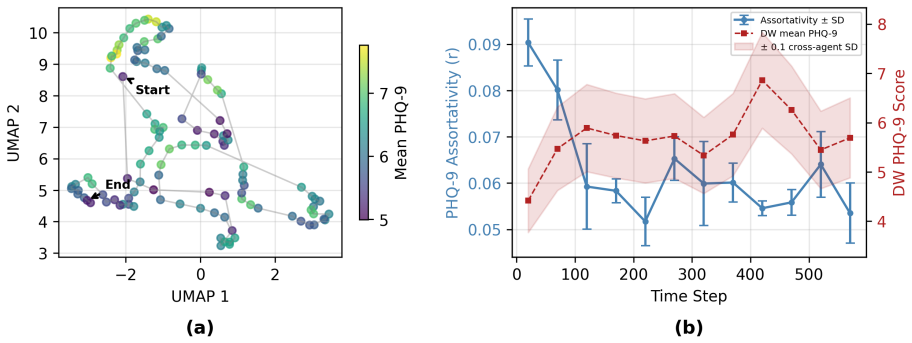

Fig. 3. Network (Mean) Degree $k = 9$: (a) UMAP visualization of MentalBERT embeddings illustrating semantic drift over time (window size = 35, shift = 5), average over full population of agents, with semantic states colored by mean PHQ-9 scores. (b) Assortativity and DW PHQ-9 throughout the simulation.

Figure 3 visualizes the temporal drift of depressive language in the MentalBERT embedding space for 200 agents over 600 timesteps with an average degree of 9. Each point represents the mean embedding of an agent's recent tweet window, and trajectories indicate how agents' linguistic expressions move through latent affective space over time. We observe a pattern of semantic drift: the language

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

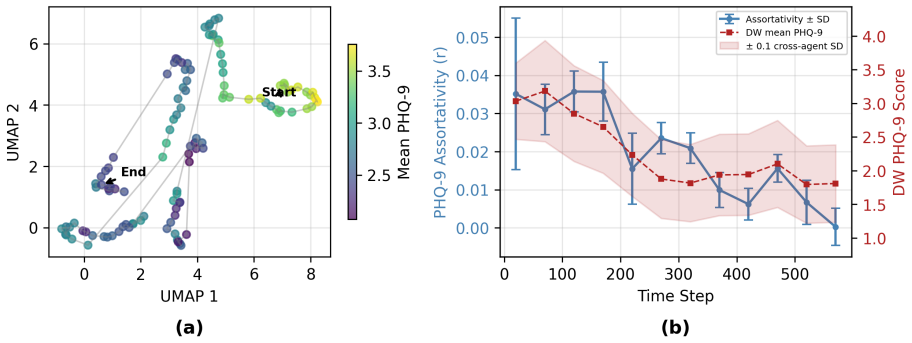

Fig. 4. Network (Mean) Degree $k = 6$: (a) UMAP visualization of MentalBERT embeddings illustrating semantic drift over time (window size = 35, shift = 5), average over full population of agents, with semantic states colored by mean PHQ-9 scores. (b) Assortativity and DW PHQ-9 throughout the simulation.

of the system moves through the latent space, and clustering of PHQ-9 values within this space shows that these latent dimensions (partly) capture the depressogenic drift. The assortativity of well-being stabilizes while the degree weighted PHQ-9 rises, suggesting the emergence of shared depressive discourse modes at the population level. Simultaneously, the sustained cross-agent variance highlights the dynamic evolution of distinct depressive clusters within the network. These trajectories demonstrate that local interaction rules and homophily-driven connectivity are sufficient to generate clustered "linguistic climates" of depression, even when agents start from heterogeneous initial conditions.

In contrast, Figure 4 demonstrates that reducing the network degree to k = 6 leads to a significant drop in depressive markers, with mean PHQ-9 scores shifting to a much lower range ($[1.5, 4.0]$) relative to the k = 9 baseline ($[5.0, 7.8]$). The drift trajectory of this simulation exhibits a less fragmented spatial distribution, showing a clearer semantic grouping of similar affective states in the latent space. This suggests that lowered peer influence induces less volatility in depressogenic language. By introducing less depressogenic reinforcement, the system is pushed into a stable, healthier climate. At the same time, variability in both PHQ-9 and assortativity persists, indicating that a subset of agents remains on peripheral or divergent paths, reflecting persistent heterogeneity in how individuals adopt or resist depressive language despite exposure to similar content.

Together, these simulations suggest a structural tipping point: increased degree amplifies the local reinforcement of depressive states, driving the emergence of depressive clusters.

4 Conclusion

In this study, we presented a novel computational framework to investigate how language entrainment facilitates the spread of distorted, depressive content. As a

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

foundational model, it successfully captures the temporal dynamics of linguistic alignment. Using advanced text embeddings, the system visualizes the natural heterogeneity of a population—highlighting both the adoption of and resistance to depressogenic discourse. Aggregating the semantic state in PHQ-9 values reveals the dynamic formation of both depressogenic and protective clustering.

Despite these contributions, several limitations need to be acknowledged. First, the PHQ-9 based well-being estimates and MentalBERT embeddings provide proxy measures of depressive symptomatology; they capture linguistic correlates rather than clinically verified diagnoses, and may therefore introduce systematic biases into inferred contagion dynamics. Second, the current implementation focuses on a limited set of network sizes and parameter regimes, leaving open how robust the observed entrainment patterns are to changes in scale, connectivity, or exogenous shocks such as moderation policies or platform design changes.

Future work will address these limitations in several directions. On the modeling side, we plan to incorporate temporal variability in susceptibility (and resilience). Methodologically, extending the framework to multi-run, large-scale experiments with systematic variation of network topology (e.g., highly modular, small-world, or scale-free structures) will enable sensitivity and uncertainty analyses of language diffusion pathways. In terms of measurement, integrating additional mental-health-specific embedding models, complementary lexicon-based indicators, and, where possible, calibration against anonymized observational data will improve validity and interpretability. Finally, the framework can be leveraged as a testbed for intervention design mitigating the spread of depressogenic content (such as content flagging) in online environments.

Disclosure of Interest. Johan Bollen is a member of the programme committee for the CompSy track of the International Conference on Computational Science (ICCS). The other authors have no competing interests to declare.

References

1. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023)
2. Ashery, A.F., Aiello, L.M., Baronchelli, A.: Emergent social conventions and collective bias in llm populations. Science Advances **11**(20), eadu9368 (2025)
3. Bathina, K.C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L.A., Bollen, J.: Individuals with depression express more distorted thinking on social media. Nature human behaviour **5**(4), 458–466 (2021)
4. Bathina, K.C., ten Thij, M., Valdez, D., Rutter, L.A., Bollen, J.: Declining well-being during the COVID-19 pandemic reveals US social inequities. PLOS ONE **16**(7), e0254114 (2021). https://doi.org/10.1371/journal.pone.0254114
5. Chan, X., Wang, X., Yu, D., Mi, H., Yu, D.: Scaling synthetic data creation with 1,000,000,000 personas (2024)
6. Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., Potts, C.: No country for old members: User lifecycle and linguistic change in online communities.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29

In: Proceedings of the 22nd international conference on World Wide Web. pp. 307–318 (2013)
7. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models (2024)
8. Durandard, N., Dhawan, S., Poibeau, T.: Language style matching in large language models. In: Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 620–636 (2025)
9. Edinger, J.D., Wohlgemuth, W.K., Radtke, R.A., Marsh, G.R., Quillian, W.C.: Cognitive behavioral therapy for treatment of chronic primary insomnia: a randomized controlled trial. JAMA **285**(14), 1856–1864 (2001). https://doi.org/10.1001/jama.285.14.1856
10. Hasan, E., Epping, G., Lorenzo-Luaces, L., Bollen, J., Trueblood, J.S.: One-shot intervention reduces online engagement with distorted content. PNAS nexus **4**(3), pgaf068 (2025)
11. Hu, Z.: Research on the impact of social media algorithmic on user decision-making: Focus on algorithmic transparent and ethical design. Applied and Computational Engineering **174**, 18–22 (07 2025). https://doi.org/10.54254/2755-2721/2025.PO24665
12. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., Cambria, E.: Mentalbert: Publicly available pretrained language models for mental healthcare. In: Proceedings of the 13th Language Resources and Evaluation Conference. pp. 7184–7190 (2022)
13. Lan, X., Han, Z., Cheng, Y., Sheng, L., Feng, J., Gao, C., Li, Y.: Depression detection on social media with large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 2155–2171 (2025)
14. Moukalled, S.H., Bickham, D.S., Rich, M.: Examining the associations between online interactions and momentary affect in depressed adolescents. Frontiers in Human Dynamics **3**, 624727 (2021)
15. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). pp. 1–22. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3586183.3606763
16. Rosen, Z.P.: A bert's eye view: A big data framework for assessing language convergence and accommodation. Journal of Language and Social Psychology **42**(1), 60–81 (2023)
17. Stronks, K., Snijder, M.B., Peters, R.J.G., Prins, M., Schene, A.H., Zwinderman, K.A.H.: Unravelling the impact of ethnicity on health in Europe: the HELIUS study. BMC Public Health **13**(1), 402 (2013). https://doi.org/10.1186/1471-2458-13-402
18. Talaga, S., Nowak, A.: Homophily as a process generating social networks: insights from social distance attachment model. arXiv preprint arXiv:1907.07055 (2019)
19. Whitty, M.T., Doherty, S.: Enhancing mis- and disinformation detection and understanding its influence: leveraging communication accommodation theory and information manipulation theory. Behaviour & Information Technology **0**(0), 1–20 (2026). https://doi.org/10.1080/0144929X.2026.2635512
20. Ye, X., Gao, H.: Distress disclosure on social media and depressive symptoms among college students: the roles of social comparison and gender. Frontiers in Psychology **16**, 1520066 (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_29