Computational Social Science Approach to Online Community Dynamics: Discovering Collective Identity Formation through Language Models

Henrik Powell
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
contacthenrik@ku.edu

Terry L. Becker
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
helloterry@uc.edu

Surieaj Krishnean
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
surajkrishnan@ucf.edu

Abstract

Online communities function as dynamic socio-technical systems in which collective identity continuously emerges through discourse, shared practices, and evolving norms. Understanding how collective identity forms and transforms within these digitally mediated spaces remains a central challenge for both social theory and platform design. This paper presents a computational social science framework that leverages large-scale language models to discover and trace collective identity formation across diverse online communities. We articulate a systems-level architecture that integrates natural language processing pipelines with theoretical constructs from social identity theory, enabling the extraction of community-level narratives, value structures, and boundary-making language. The discussion emphasizes structural trade-offs in designing such discovery systems, including tensions between batch and stream processing, model interpretability, computational scalability, and privacy preservation. We examine governance implications arising from algorithmic surfacing of collective identity, addressing fairness, representation, and the risk of essentializing community voices. Further, we analyze infrastructure requirements for sustainable deployment and the robustness of identity models under linguistic drift and platform evolution. Through cross-domain case illustrations spanning open-source developer communities, health support forums, and fan subcultures, we highlight how language model outputs can inform platform moderation policies, community health diagnostics, and inclusive design. The paper contributes a holistic, systems-oriented perspective on computationally studying collective identity, foregrounding the interplay between technical architecture, social theory, and responsible deployment in large-scale online environments.

Keywords

computational social science, online communities, collective identity, language models, natural language processing, socio-technical systems, community governance, fairness, infrastructure.

1. Introduction

Online communities have become foundational structures of contemporary social life, hosting everything from professional collaboration and civic mobilization to emotional support and cultural production. Within these digitally mediated collectives, participants continuously negotiate who "we" are, what "we" value, and how "we" differ from others. This ongoing process of collective identity formation shapes community cohesion, norm enforcement, knowledge sharing, and conflict dynamics [1]. Despite decades of qualitative and quantitative research, capturing the multifaceted, emergent nature of collective identity at scale remains an open problem. The rise of computational social science offers new opportunities to model such complex social phenomena using large-scale digital trace data and machine learning techniques [1]. In particular, pretrained language models, ranging from early distributional semantic models to contemporary transformer-based architectures, have demonstrated a remarkable capacity to encode relational meaning, discourse patterns, and cultural associations embedded in text. This paper proposes a systems-oriented framework that harnesses these language models to systematically discover and monitor collective identity formation in online communities, bridging micro-level linguistic choices with macro-level community structures.

The study of collective identity has deep roots in social psychology and sociology, where seminal work established that individuals derive part of their self-concept from perceived membership in social groups, engaging in social comparison and boundary maintenance [2]. Translating these theoretical insights into computational models requires more than simply applying topic models or clustering algorithms to forum posts; it demands a careful integration of social theory with data engineering, model architecture, and deployment considerations. The computational infrastructure must support the ingestion of heterogeneous, high-velocity text streams while respecting ethical boundaries and platform constraints. Language models, when employed as analytic lenses, reveal how shared lexicons, narrative frames, and rhetorical strategies coalesce into durable identity signals. However, the deployment of such systems introduces significant structural trade-offs. Real-time identity monitoring competes with the latency and cost of large model inference; model interpretability vies with predictive performance; and the need to generalize across communities conflicts with the imperative to honor local contextual specificities.

This paper addresses these tensions from an interdisciplinary systems perspective. We argue that discovering collective identity through language models is not merely a modeling exercise but a socio-technical design problem that intertwines computational architecture, governance policy, and sustainability planning. The subsequent sections develop this argument by first revisiting relevant theoretical and technical foundations, then outlining a modular system architecture, and finally examining governance, fairness, and robustness challenges. Throughout, we emphasize system-level discussions rather than algorithmic minutiae, focusing on structural choices that shape the viability and responsibility of deploying such systems at scale.

2. Theoretical Foundations and Related Work

Collective identity in online settings has been studied through diverse lenses, including netnography, discourse analysis, and large-scale quantitative content analysis. Early ethnographic approaches adapted to digital contexts provided rich accounts of how communities develop shared languages, rituals, and insider humor that together constitute a "we-ness" [4]. These qualitative insights underscored the importance of boundary work, in which groups differentiate insiders from outsiders through specialized vocabulary and

rhetorical norms. Simultaneously, computational methods for analyzing textual data have rapidly evolved. Latent Dirichlet Allocation (LDA) introduced a generative probabilistic model that represents documents as mixtures of latent topics, enabling researchers to uncover thematic structures in large corpora without predefined categories [5]. While powerful, topic models primarily capture coarse thematic content and struggle to encode the nuanced relational meanings that underpin identity.

The neural turn in natural language processing brought distributional semantic models such as Word2Vec, which learn continuous vector representations of words from their co-occurrence contexts [6]. These embeddings captured semantic and syntactic regularities, making it possible to measure cultural associations and stereotypes encoded in language. The subsequent development of contextualized language models, most notably BERT, represented a paradigm shift by learning dynamic word representations that depend on surrounding textual context [7]. Such models opened new frontiers for studying how meaning shifts across communities and over time, a crucial capability for tracking identity evolution. Diachronic word embedding techniques further extended this capacity by aligning embedding spaces across time slices, revealing statistical regularities in semantic change that can signal shifting collective concerns and boundaries [8].

Computational sociolinguistics has emerged as a vibrant interdisciplinary field that applies such models to understand social dimensions of language variation, including identity expression, group membership signaling, and power relations [9]. These studies demonstrate that linguistic patterns such as pronoun use, formality, and affective expression serve as reliable markers of community identification. However, bridging individual-level linguistic choices to emergent, system-level collective identity requires more than aggregating lexical features; it demands architectures that can synthesize signals across multiple levels of analysis, from individual posts to entire discussion threads to community-wide narrative arcs. The present work builds on these foundations by proposing a holistic system design that intentionally couples language model outputs with theoretical frameworks of social identity, drawing inspiration from recent explorations of self-perception modeling [10] and from studies correlating linguistic variation with salient social categories such as gender [11].

3. A Computational Systems Architecture for Identity Discovery

Designing a computational architecture for discovering collective identity necessitates modular decomposition that balances analytical depth with operational scalability. At the highest level, the proposed system comprises four interconnected layers: data ingestion and preprocessing, language model inference, identity signal aggregation, and interpretive feedback. The ingestion layer must accommodate the heterogeneous nature of online community data, which includes not only textual posts and comments but also metadata such as timestamps, user identifiers, voting scores, and platform-specific structural features. Choices at this layer incur immediate trade-offs between batch-oriented processing, which enables comprehensive historical analysis at lower infrastructure cost, and stream processing, which supports near-real-time identity monitoring essential for responsive community governance. Hybrid lambda architectures that combine batch and speed layers represent a pragmatic compromise, though they introduce operational complexity in maintaining consistency across processing paths.

The language model inference layer forms the analytical core, where pretrained transformer models are applied to encode posts into dense vector representations. Here, system designers face pivotal decisions regarding model size, fine-tuning strategy, and embedding extraction

methodology. Large general-purpose models offer rich semantic representations but demand substantial computational resources, raising sustainability concerns that become acute when scaled to platforms hosting millions of daily interactions. Distilled or domain-adapted models can mitigate resource consumption, yet they risk losing the breadth of cultural knowledge that makes large models effective for cross-community comparison. Additionally, the choice between extracting static sentence embeddings for clustering and performing more complex generative tasks for narrative summarization directly influences downstream interpretability. Architecturally, containerized microservices that orchestrate model inference allow elastic scaling, while caching mechanisms for frequently encountered textual patterns can further improve throughput without sacrificing analytical fidelity.

The aggregation layer synthesizes individual text representations into community-level identity constructs. This process involves clustering semantically similar posts, tracking the temporal dynamics of cluster prominence, and identifying linguistic features that distinguish one community from others. Crucially, aggregation must be theory-guided; raw similarity matrices alone do not constitute collective identity. The system incorporates social identity theory constructs by searching for clusters of language that correspond to shared values, common goals, and intergroup differentiation. For example, in open-source software communities, recurring linguistic patterns around meritocracy, contribution ethics, and technical jargon can be mapped onto communal identity dimensions. The aggregation layer outputs structured representations of collective identity, such as multidimensional vectors capturing community ethos, boundary markers, and narrative themes, which can then be consumed by visualization dashboards or governance application programming interfaces.

The interpretive feedback layer closes the loop by enabling human analysts and community moderators to inspect, validate, and refine machine-discovered identity signals. This socio-technical integration acknowledges that collective identity is ultimately a human-interpreted construct and that fully automated discovery risks perpetuating reductive or biased representations. By providing interfaces that trace aggregated identity markers back to exemplar posts and allow community members to contest or endorse algorithmic characterizations, the system fosters participatory sensemaking. The architectural implication is that the pipeline must maintain rich provenance metadata, linking each high-level identity signal to its supporting textual evidence. This design principle not only supports interpretability but also strengthens accountability when identity discoveries inform moderation actions or policy recommendations.

4. Language Model-Driven Extraction of Collective Identity

The application of language models to collective identity formation operates at the intersection of representation learning and sociolinguistic theory. Unlike individual identity, which can be probed through self-description, collective identity manifests emergently in the aggregate linguistic patterns of many interacting individuals. Language models, particularly those based on transformer architectures pre-trained on massive corpora, serve as powerful tools for surfacing these latent patterns because they encode distributional knowledge about word usage, discourse structure, and cultural associations. When fine-tuned on community-specific data, these models can adapt their internal representations to capture local meanings that constitute in-group language. Recent studies have demonstrated that language model representations can capture aspects of how people conceptualize themselves, providing a foundation for extending such insights to group-level self-conceptions [10]. Building on these

findings, systematic comparison of embedding spaces derived from different communities can reveal divergent semantic organizations that correspond to distinct collective identities.

A particularly valuable analytical strategy involves probing the geometry of the embedding space to identify axes of variation that align with theoretical identity dimensions. For example, the distinction between inclusive "we" and exclusive "they" can be operationalized by examining how community members linguistically construct in-group and out-group categories. Language models enable large-scale analysis of such constructions by identifying the typical contexts in which first-person plural pronouns appear and the semantic properties of the predicates associated with them. Extending this line of inquiry, lexical variation studies grounded in social categories, such as gender, have shown that seemingly subtle stylistic differences carry rich identity information [11]. In online health support communities, for instance, distinct linguistic profiles emerge around illness-centric versus recovery-centric identities, which can be automatically detected through clustering of model-derived post embeddings. The evolution of these profiles over time offers a computational window into how identity solidifies or fragments in response to external events or internal conflicts.

Cross-community comparison represents another powerful modality. By embedding posts from multiple communities into a shared representational space, researchers can measure inter-community distances and identify linguistic features that are uniquely diagnostic of particular collectives. A systems design that supports such multi-community analysis must address metadata standardization, sampling strategies that avoid temporal confounding, and statistical methods for significance testing of observed differences. Importantly, drawing on diachronic embedding methodologies [8], the system can detect semantic drift in key identity terms, flagging moments when a community's core vocabulary undergoes redefinition. Such signals serve as early warnings of identity contestation or transformation, providing valuable intelligence for community stewards. In fan subcultures, for example, shifts in the meaning of genre-specific terminology can indicate tensions between long-standing members and newcomers, an identity dynamic that computational methods can surface well before it erupts into overt conflict.

5. Infrastructure and Deployment Considerations

Deploying a language model-based identity discovery system in production environments introduces infrastructure demands that extend well beyond model training. Data pipelines must reliably ingest, clean, and store vast volumes of user-generated content while complying with platform terms of service and jurisdictional data protection regulations. Message queuing systems inspired by distributed log technologies have become essential for decoupling data producers from consumers, enabling resilient stream processing [12]. These infrastructure components must be designed to handle spiky workloads characteristic of online platforms, where breaking news or viral events can cause sudden spikes in community activity. Auto-scaling compute clusters provisioned through cloud infrastructure provide the elasticity needed to absorb such bursts without maintaining costly over-provisioned capacity during quiescent periods [13]. The system must also implement strict data retention and anonymization policies; identity discovery often operates on pseudonymized data, but re-identification risks remain a persistent challenge that must be mitigated through differential privacy techniques or on-device processing where feasible.

The computational cost of repeatedly applying large language models to every incoming post can quickly become prohibitive. System architects must therefore implement tiered analysis strategies: lightweight models perform initial filtering and triage, while more expensive deep

semantic analysis is reserved for content deemed highly relevant to identity processes based on keyword triggers, user engagement signals, or temporal proximity to known community events. Model caching and embedding reuse strategies further reduce computational overhead, as many posts in online communities exhibit formulaic structures such as greetings, routine questions, or platform-mandated templates. Beyond cost, the carbon footprint of large-scale language model inference has emerged as a significant sustainability concern that cannot be divorced from system design [19]. Architects bear responsibility for incorporating energy-aware scheduling, leveraging efficient hardware accelerators, and evaluating whether marginal gains in identity detection accuracy justify disproportionate energy expenditure.

Robustness against distributional shift constitutes another critical deployment challenge. Language in online communities evolves rapidly; new slang, novel acronyms, and shifted semantic associations can render static language models increasingly inaccurate over time. Concept drift literature provides frameworks for monitoring model performance degradation and triggering retraining cycles [17]. A sustainable deployment plan must therefore include continuous integration pipelines that regularly evaluate model outputs against human-annotated holdout sets and automatically retrain or fine-tune models when drift exceeds acceptable thresholds. This requirement introduces feedback loops that reinforce the socio-technical nature of the system: community members themselves contribute to the evolving language distributions that the system must track, creating a reflexive relationship that demands careful governance to avoid runaway feedback where the system's outputs unduly influence the very linguistic behaviors it measures.

6. Governance, Fairness, and Policy Implications

The capability to computationally discover collective identity carries profound governance implications for online platforms. When an algorithmic system surfaces and labels community identity characteristics, it engages in a form of representational power that can shape how communities are perceived by platform operators, advertisers, researchers, and the communities themselves. Fairness concerns emerge along multiple axes. First, language models pre-trained on dominant cultural datasets may systematically misrepresent or fail to capture the identity expressions of marginalized or non-Western communities, perpetuating representational harms [18]. Mitigating such biases requires careful curation of training corpora and community-specific model adaptation, as well as adversarial testing for stereotype amplification. Second, algorithmic surfacing of collective identity can inadvertently essentialize fluid, contested, or intersectional community memberships into fixed categories, risking the erasure of internal diversity. System designers must build skepticism into the architecture by exposing uncertainty estimates and supporting multi-faceted identity representations rather than singular labels.

Content moderation practices intersect with identity discovery in complex ways. Platforms increasingly rely on automated tools to detect hate speech, harassment, and coordinated inauthentic behavior, tasks that often depend on understanding community norms and identity cues. However, using identity models to inform moderation risks creating feedback loops where majority identity norms are algorithmically enforced, stifling legitimate dissent or subcultural innovation. Gillespie's analysis of platform custodianship underscores how moderation decisions are never purely technical but rather instantiate specific value judgments about acceptable speech and community boundaries [15]. An identity-aware moderation system must therefore be transparent about its operational thresholds and subject to adversarial appeal processes that allow communities to contest machine-generated

characterizations. From a policy perspective, the design of such systems should be guided by procedural fairness principles, ensuring that affected communities have meaningful opportunities to participate in defining how their collective identity is represented and acted upon.

The technical debt accumulated by machine learning systems further complicates long-term governance [16]. Identity discovery pipelines consist of complex interdependencies among data ingestion, feature computation, model inference, and aggregation logic. Changes in any upstream component can propagate in unanticipated ways, altering identity outputs in ways that are difficult to audit. Robust governance requires versioned data lineage tracking and automated regression testing that compares identity outputs over time to detect anomalies introduced by system modifications. Furthermore, the ethical frameworks governing human subjects research provide instructive parallels for computational identity discovery. The emerging ethics divide around big data research highlights that traditional informed consent models are ill-suited to large-scale observational studies where individuals may be unaware their data contributes to identity profiling [20]. Infrastructure designs that embed privacy-preserving computations and community-level opt-out mechanisms can partially address these concerns while maintaining analytic utility.

7. Sustainability and Long-Term Robustness

Sustaining a computational identity discovery system over years, rather than as a one-off academic project, demands attention to both technical and organizational robustness. On the technical side, model staleness caused by linguistic drift remains a primary threat. As communities evolve, their lexicon, communication norms, and identity narratives change, degrading the fidelity of static models. While periodic retraining addresses this drift, it introduces a tension with reproducibility: research findings derived from one model version may not replicate under an updated model, complicating longitudinal studies. System architectures should therefore support snapshot archiving, allowing researchers to reconstruct earlier model states and contextualize findings against the specific analytical lens through which they were generated. Versioned data lakes and model registries become essential infrastructure components for maintaining scientific integrity.

Organizational sustainability hinges on embedding the identity discovery system within broader community management workflows. Tools that generate actionable insights for moderators, community managers, and health analysts are more likely to receive sustained investment than purely research-oriented deployments. For example, identity diagnostics that detect emerging factionalism or declining common ground can inform proactive interventions such as facilitated dialogues or restructuring of discussion spaces. However, instrumentalizing identity in this way raises ethical questions about the manipulation of community dynamics. The boundary between supporting community health and engineering social outcomes is porous and must be navigated through transparent governance structures that involve community representatives in decision-making processes.

Energy and resource consumption also intersect with sustainability mandates. The environmental impact of training and serving large language models has come under increasing scrutiny, prompting research into more efficient architectures and inference strategies [19]. System designers can reduce footprints by applying model distillation, activating only task-specific adapter modules, and scheduling compute-intensive jobs during periods of high renewable energy availability on the grid. These considerations, while seemingly orthogonal to identity theory, are integral to responsible system engineering in an

era where the carbon implications of AI infrastructure are no longer externalizable. A truly sustainable identity discovery system is one that balances analytical power with resource parsimony, guided by continuous cost-benefit re-evaluation of the infrastructure's footprint relative to the social value it generates.

8. Conclusion

This paper has presented a comprehensive, systems-oriented examination of how computational social science, instantiated through language model-driven architectures, can illuminate the processes of collective identity formation in online communities. By moving beyond narrow algorithmic treatments and instead foregrounding the interplay among theoretical grounding, system architecture, governance, and sustainability, we have argued that discovering collective identity is fundamentally a socio-technical design challenge. The proposed framework demonstrates how modular architectures can integrate ingestion, inference, aggregation, and feedback layers to surface identity signals while grappling with trade-offs between real-time responsiveness and computational cost, between model fidelity and interpretability, and between analytical insight and representational fairness. Structural choices made at each layer reverberate across the system, influencing who gets to define community identity and with what consequences.

Governance and policy implications arise as first-class design constraints rather than afterthoughts. The power to algorithmically characterize collective identity carries responsibilities to mitigate bias, preserve internal diversity, and enable contested interpretations. Embedding participatory feedback loops, provenance tracking, and community opt-out mechanisms into the system architecture represents a step toward more accountable computational social science. Simultaneously, the long-term viability of such systems depends on addressing model drift, technical debt, and energy consumption through disciplined engineering practices and institutional commitment. As online communities continue to serve as critical sites for social connection, political discourse, and cultural production, understanding the computational underpinnings of their collective identities becomes increasingly urgent. The interdisciplinary approach advanced here provides a blueprint for building systems that are not only technically sophisticated but also socially reflexive and ethically grounded, capable of supporting vibrant digital publics in an era of planetary-scale online interaction.

References

1. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. Science, 323(5915), 721–723.

2. Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), Psychology of intergroup relations (pp. 7–24). Nelson-Hall.

3. boyd, d., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210–230.

4. Kozinets, R. V. (2010). Netnography: Doing ethnographic research online. Sage Publications.

5. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

8. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 1489–1501).

9. Nguyen, D., Dogruoz, A. S., Rose, C. P., & de Jong, F. (2020). Computational sociolinguistics: A survey. Computational Linguistics, 46(3), 567–639.

10. Solanki, D., Hsu, H. M., Zhao, O., Zhang, R., Bi, W., & Kannan, R. (2020, July). The way we think about ourselves. In International Conference on Human-Computer Interaction (pp. 276-285). Cham: Springer International Publishing.

11. Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics, 18(2), 135–160.

12. Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system for log processing. In Proceedings of the NetDB 2011 Workshop.

13. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58.

14. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

15. Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

16. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems 28 (pp. 2503–2511).

17. Zliobaite, I. (2010). Change with delayed labeling: When is it detectable? In 2010 IEEE International Conference on Data Mining Workshops (pp. 843–850).

18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

19. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

20. Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. Big Data & Society, 3(1), 1–14.

21.Zhou, D. (2026, May). A Code Visualization Graph-Based Method for Vulnerability Severity Assessment Using a Multi-Scale Feature Fusion Network. In 2026 3rd

International Conference on Image Processing and Artificial Intelligence (ICIPAI) (pp. 306-309). IEEE.