University of Kent

Kent Academic Repository

Hill, Mark J., Miller, Vincent and Moreira, Tiago (2026) *Discursive Signatures: A Computational Method for Mapping How Meaning Varies Across Communities*. In: Anthology of Computers and the Humanities. Anthology of Computers and the Humanities Volume 6. 6. pp. 46-56. Association for Computers and the Humanities, Austin, Texas, USA

Downloaded from
https://kar.kent.ac.uk/117674/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.63744/dVXq85XJAhwJ

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Discursive Signatures: A Computational Method for Mapping How Meaning Varies Across Communities

Mark J. Hill[1], Vince Miller[2], and Tiago Moreira[3]

[1] Department of Digital Humanities, King's College London, London, UK
[2] School of Social Policy, Sociology and Social Research, University of Kent, Canterbury, UK
[3] Department of Sociology, Durham University, Durham, UK

Abstract

This paper presents a computational method for studying how the meaning of cultural artefacts varies across online communities. Rather than treating topic model outputs as thematic labels, we propose using full topic-proportion vectors as discursive signatures that capture meaning at both the level of individual expression and the community contexts that shape it. By aggregating these signatures at the community level, validating groupings with PERMANOVA, applying hierarchical clustering, and projecting the resulting space via PCA, we produce a topology of discourse: a spatial representation of how meaning-making varies systematically across contexts. We demonstrate the method through a case study of mental health technology discourse on Reddit (289,039 posts across 445 subreddits), showing how the same technologies are conceptualised differently depending on community context. The method is designed to be transferable to any domain where researchers seek to understand how technologies, practices, or concepts are variably constituted across online spaces.

Keywords: computational discourse analysis, cultural analytics, social media, community-level modelling, discourse topology, mental health technology

1 Introduction

A persistent challenge in computational humanities and cultural analytics is reconciling scale with interpretive sensitivity. Computational methods achieve breadth (processing hundreds of thousands of documents across dozens of communities) but often flatten contextual variation that constitutes meaning. Ethnographic and discourse-analytic approaches preserve content but cannot systematically compare across many communities simultaneously. This tension is particularly acute when studying how cultural artefacts or practices are understood differently across social contexts: the same object may carry fundamentally different meanings depending on who discusses it, where, and within what norms.

This paper addresses that challenge by presenting a computational method for mapping how meaning varies across online communities. Drawing on Computational Grounded Theory [14], which integrates computational pattern detection with interpretive analysis, we develop a pipeline that moves from corpus construction through topic modelling to a geometric representation of discourse. The key methodological move is treating topic model outputs not as categorical labels for documents but as discursive signatures in which their full topic-proportions (θ-vectors) capture how a community characteristically talks about a phenomenon, including registeral and interactional features alongside substantive content [3; 4]. By aggregating, clustering, and projecting

Mark J. Hill, Vince Miller, and Tiago Moreira. "Discursive Signatures: A Computational Method for Mapping How Meaning Varies Across Communities." In: *Computational Cultural Science Workshop 2026*, ed. by Suzanne Mpouli, Jean-Baptiste Camps, Florian Cafiero, and Chahan Vidal-Gorène. Vol. 6. Anthology of Computers and the Humanities. 2026, 46–56. https://doi.org/10.63744/dVXq85XJAhwJ.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

these signatures, we produce a topology of discourse in which spatial proximity encodes discursive similarity and distance encodes divergence in how meaning is framed.[1]

We demonstrate the method through a case study of mental health technology (MHT) discourse on Reddit. This is a domain where the same app, platform, or artificial intelligence (AI) tool is discussed across communities organised around different conditions, identities, and needs. While this application is substantively important, it is not the paper's primary contribution. Rather, the method is a transferable framework for any domain where researchers wish to study how cultural objects are variably constructed across spaces. The pipeline's components (keyword-based corpus construction, subreddit categorisation, mixed-membership topic modelling, community-level aggregation, hierarchical clustering, and PCA) are individually presented as constituent parts of an analytical workflow for studying contextual meaning-making.

Theoretically, the method treats meaning as a situated practice rather than stable reference. In this case, what a technology is (its function, its value, the kind of care it enables) is not fixed by its design but emerges through the practices of communities [20]. While this position has roots in several traditions, what unites them is an understanding that meaning is not individually-held, but emerges from, and is embedded in, social exchanges [5; 9; 23].

This foregrounds why we examine discourse at the community level, why we treat topic distributions as encoding practice rather than merely content, and why the interpretive steps (topic labelling, community categorisation, close reading of high-probability documents) are analytically central. Figure 1 outlines this integration. The content dimension (documents, topics, probability distributions) captures textual meaning; the context dimension (subreddits, subject domains, community metadata) captures the social conditions of meaning making. Similarly, the move from individual posts (micro) to topics (macro) and individual communities (micro) to subject-domain communities (macro) is not merely an aggregation procedure, but a shift in the unit of analysis. Meaning that is not legible at the level of the individual utterance becomes visible when the characteristic patterns of communities are examined as a whole.

The remainder of this paper describes each stage of the pipeline, the decisions and trade-offs involved, and what the method reveals when applied to MHT discourse on Reddit. We conclude by discussing transferability, limitations, and implications for computational approaches to cultural analytics.

2 Data and Corpus Construction

The method begins with constructing a corpus that captures both content (what is said) and context (where it is said). We describe the approach using our case study of MHT discourse, though the principles generalise.

2.1 Source Selection: Reddit as Structured Discourse

Reddit's architecture makes it particularly suited to this method. Discussions are organised into topic-oriented forums (subreddits) with distinct norms and moderation practices [19]. Subreddits themselves are often made up of users with shared identities, conditions, or interests. Each subreddit carries metadata (title, sidebar description, rules) that allow researchers to characterise its social function independently of the text it contains. This is what the method requires: a platform where documents are naturally nested within identifiable communities, providing both textual content and contextual metadata.

There are additional benefits, however. Platform affordances, including anonymity and 'karma,' facilitate highly detailed and personal discussions [6]. It is for this reason that Reddit has

[1] While this paper presents a methodological pipeline, it is part of a broader research project examining mental health technology[11; 13].

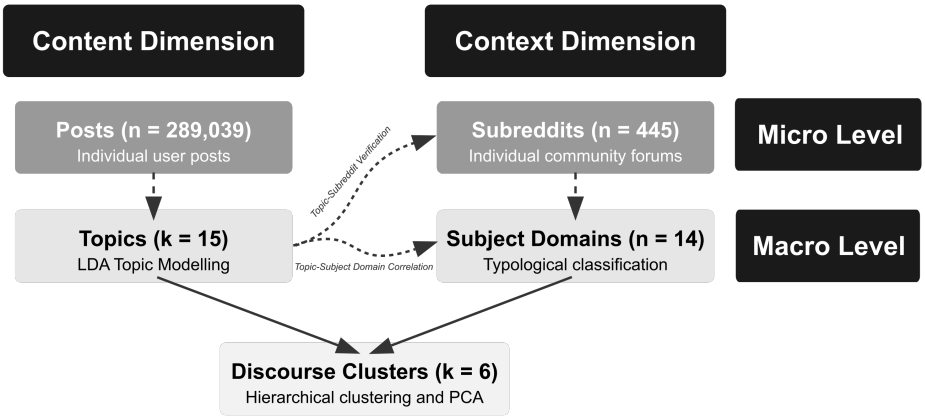


Figure 1: Analytical pipeline showing how content (posts, topics) and context (subreddits, subject domains) dimensions are integrated through topic modelling and hierarchical clustering to produce discourse clusters.

been used extensively to study topics ranging from identity construction to mental health [1; 8; 10; 15; 16; 17; 24].

2.2 Keyword Design and Subreddit Selection

Corpus construction proceeds through keyword-based search followed by community-level filtering. For MHT, we developed a search vocabulary of over 140 terms spanning five categories: (1) named platforms (e.g., Daylio, Replika, BetterHelp); (2) modality descriptors (e.g., 'online therapy', 'therapy app'); (3) digitally mediated therapeutic techniques (e.g., 'CBT app', 'meditation app'); (4) AI-specific terms (e.g., 'AI therapist', 'therapy chatbot'); and (5) evaluative language (e.g., 'therapy replacement').[2] This multi-category approach aimed to capture discourse that references specific technologies, generic categories, and evaluative framings.

From initial results, subreddits were identified and manually reviewed, with communities reflecting non-target contexts (e.g., AI discussion oriented toward gaming) excluded. 445 subreddits were kept and classified into one of fourteen subject domains based on subreddit title, sidebar descriptions, and qualitative review of posts to assess perceived social function. The aim was to identify why users would choose to post in a given subreddit. This categorisation is necessary: it provides the community-level metadata against which computational patterns are subsequently validated. Table 1 provides an overview.

2.3 Preprocessing

After removing bot-generated posts, spam, non-English content, and posts consisting entirely of search terms, the corpus comprises 289,039 documents from 218,656 unique users, spanning July 2008 to October 2025.

Text pre-processing involved three decisions warranting discussion. First, statistically significant collocations were identified and merged (e.g., 'ai_therapy') to preserve multi-word concepts that single-token approaches would dissolve. Second, lemmatisation was applied to reduce inflectional variation. Third, stopword removal was conducted with theoretically motivated exceptions. Standard stopword lists remove function words, but in discourse about technology use several categories carry analytic weight: pronouns (distinguishing self-disclosure from advice-giving), com-

[2] Terms were expanded to account for spelling variations, hyphenation differences, and informal abbreviations.

Subreddit Domain	Posts	Subreddits
Mental health conditions	66,396	96
Technology & AI	57,658	56
Self-improvement	51,584	70
Knowledge-sharing & Q&A	25,773	17
Relationships & family	25,223	54
Professional mental healthcare	20,900	18
Gender & sexual identity	18,978	46
Neurodivergence	16,953	30
Physical & medical health	13,044	58
Emotional outlet	7,811	15
Identity (cultural, generational)	6,916	20
Substance use & addiction	6,661	15
Abuse & trauma support	4,301	22
Medication & pharmacology	2,301	13

Table 1: Post and subreddit counts per subject domain.

mon verbs describing engagement ('use', 'try', 'got'), and temporal markers structuring narratives ('during', 'after', 'before'). These were retained. Such decisions are consequential for downstream topics [7] and therefore must be made with reference to a specific study.

3 Topic Modelling as Discursive Fingerprinting

3.1 Model Specification

We used the Structural Topic Model (STM) [18] with Latent Dirichlet Allocation (LDA) initialisation and no covariates. While STM supports the inclusion of document-level metadata as covariates, we excluded these. The reason is methodological: we aimed to test whether community-level discursive patterns emerge from text alone, without being conditioned on metadata. We do use domain labels downstream for both interpretation and as a validity check (PERMANOVA). This allows us to test convergence between manually assigned community categories and emergent topic distributions rather than assume it, and provide stronger evidence that the patterns are discursive rather than artefacts of the model.

3.2 Model Selection

Model selection followed systematic evaluation across multiple K (the number of topics), comparing held-out likelihood [21], semantic coherence, and exclusivity [12; 18]. These metrics often trade off against each other: higher K increases exclusivity (topics become more distinct) but reduces coherence and risks fragmentation. Lower K produces broader, more coherent topics but may conflate substantively distinct discourses.

We evaluated multiple models across multiple $K = \{5...30\}$. For each, we assessed: (1) held-out likelihood on a reserved document set, measuring predictive performance; (2) mean semantic coherence across topics, measuring whether high-probability words within topics co-occur in documents; and (3) mean exclusivity, measuring whether high-probability words are distinctive to their assigned topic. We selected $K = 15$ as offering the best balance: interpretable, substantively distinct topics without excessive fragmentation, and stable performance across metrics.

3.3 Topic Labelling and Interpretation

Topic labels were assigned through analysis of high-probability terms, FREX (frequent and exclusive terms), keyness statistics, and close reading of several hundred high-θ documents per topic. This yielded three categories of topic: *substantive themes* (e.g., Professional therapy, AI companion platforms, Medication, Mood & symptom tracking); *registeral modes* (e.g., Support seeking, Support giving, Conversational register); and *structural patterns* (e.g., Automated and/or structured responses capturing templated text, such as emergency help lines).

Importantly, rather than treating the presence of registeral topics as noise we take it as analytically constitutive. In mental health discourse, the same substantive concern (e.g., therapy access) may be discussed in different communities with fundamentally different interactional modes: soliciting peer advice, providing professional guidance, or venting in distress. This framing has a theoretical basis. Following work on register as a continuous rather than categorical phenomenon [4], we treat topics as encoding not only what a text is about, but also how it engages with a subject: through what register, with what interactional mode, and from what subject position.

3.4 From Topics to Discursive Signatures

This leads to the paper's central methodological claim. Standard applications of topic models assign each document to its highest-probability topic or use topics as categorical labels. We depart from this by treating each document's full θ-vector (the complete probability distribution across all K topics) as its discursive signature. While all posts may be generally discussing a similar topic (in our case, MHT), a post dominated by 30% Support seeking, 25% Medication, 20% Mood tracking carries a different discursive character than one that is 40% AI companion and 25% Medication.

Importantly, topics do not distribute independently: some pairs co-occur frequently (e.g., Familial relationships with Abuse support), while others rarely appear together (e.g., AI companion discourse and Professional therapy). This attraction-repulsion structure is both a useful early sense-check on categorisations, and analytically informative as it reflects the discursive organisation of the space under study.

4 Community-Level Aggregation and Clustering

4.1 Aggregation

Having established document-level discursive signatures, the method moves to the community level. For each subreddit, we computed the mean θ-vector across all constituent documents, producing a 15-dimensional profile characterising that community's discourse.[3] These vectors were then further aggregated to the subject-domain level (the fourteen categories assigned during corpus construction), yielding per-domain estimates of topic distribution.

This two-stage aggregation serves distinct purposes. Subreddit-level profiles capture fine-grained community variation; domain-level profiles enable comparison of broader community types. Both are used in subsequent analysis.

4.2 Validation with PERMANOVA

To ensure the categorisation did not impose structure that the text does not support, we tested whether the assigned subject-domain categories correspond to genuinely distinct discursive profiles. To do this we applied PERMANOVA [2] to subreddit-level θ-vectors grouped by domain assignment, testing whether between-group variance exceeds within-group variance. PERMANOVA

[3] We compared mean and median and observed negligible differences in clustering structure and PCA geometry.

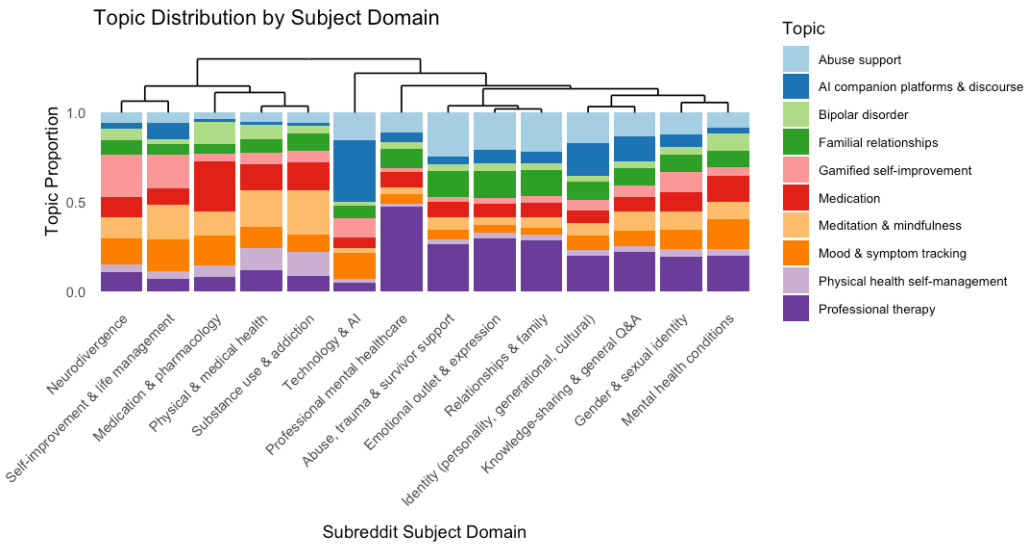


Figure 2: Topic distribution by subject domain, with hierarchical clustering dendrogram (Ward's D^2). Bars show mean topic proportions across subreddits within each domain; dendrogram reveals how domains group by discursive similarity rather than only nominal subject matter.

confirmed statistically significant differences, providing evidence that the assigned categories capture real discursive distinctions rather than arbitrary groupings. This validation step is important: it establishes that the manual categorisation and the computational topic profiles converge, allowing for subsequent analysis that integrates both.

4.3 Hierarchical Clustering

To identify higher-order organisation among community types, hierarchical clustering was performed on domain-level mean topic-proportion vectors using Euclidean distance and Ward's D^2 linkage [22]. See Figure 2.[4]

A six-cluster solution emerged that was both statistically and substantively interpretable.[5] These clusters were labelled Peer support, Therapy, Crisis, Somatic regulation, Executive scaffolding, and Algorithmic intimacy. Each represents distinct regimes of discourse, characterised by particular configurations of topics (See Table 2). Importantly, these groupings emerge from topic distributions alone and are discovered through the clustering of domain-level discursive profiles. That they align with interpretable social distinctions is evidence of the method's validity.

5 Principal Component Analysis and Discourse Topology

While hierarchical clustering reveals how community types group, it does not capture the continuous structure of variation, nor does it show how individual communities within clusters relate

[4] To improve readability topics stable across categories were dropped. These are: Conversational register, Therapeutic techniques, Support seeking, Support giving.

[5] We inspected merge heights, compared cut levels, and checked the clustering structure was stable to mean versus median aggregation. Average silhouette widths were modest and did not indicate a decisive optimum, though the six-cluster solution yielded the second-highest value among candidate cuts. We therefore retained six clusters as the best balance between quantitative separation and interpretive resolution.

Cluster	Orientation	Role of MHT	Subject domains
Peer support	Mutual aid and informal emotional support	A general resource discussed alongside broader help-seeking; communities serve as an entry point	Mental health conditions, identity, knowledge-sharing
Therapy	Professionally-oriented clinical care	Central to therapeutic relationships; clinical legitimacy anchors digital tools	Professional mental healthcare
Crisis	Acute distress and emergency support	Valued as survival infrastructure enabling access to professional help. Evaluated for access, not features or efficacy	Relationships & family, Abuse, trauma & survivor support, Emotional outlet & expression
Somatic regulation	Managing physical symptoms and medical recovery	Lay clinical instrument: logs and metrics track illness, medication effects, and recovery under clinical oversight	Medication, Physical & medical health, Substance use & addiction
Executive scaffolding	Compensating for impaired self-regulation	External structure for daily functioning when internal executive capacity feels unreliable	Neurodivergence, Self-improvement & life management
Algorithmic intimacy	Emotional relationships with AI companions	An affective platform valued for companionship (and its positive externalities), explicitly independent from clinical authority	Technology & AI

Table 2: Discourse clusters: orientation, function, and constituent subject domains.

to one another. To address this, we applied PCA to subreddit-level θ-vectors, producing a low-dimensional spatial representation of discourse.

5.1 Transformation and Dimensionality Reduction

Topic proportions are compositional data (they sum to 1 for each document), which violates the assumptions of standard PCA. We therefore applied the centred log-ratio (CLR) transformation before computing principal components. Only subreddits with at least 50 posts were included, ensuring reliable mean estimates. The first two principal components explain 64.6% of total variance, indicating strong low-dimensional latent structure.

5.2 Interpreting the Topology

PCA loadings reveal how topics relate to one another geometrically. Topics that load in the same direction co-occur across communities; topics that load in opposite directions are in tension. In our case study, this geometry proved interpretable along two axes:

PC1 (38.3% variance) separates relational exposure and need from self-governance. Topics such as Abuse support, Professional therapy, and AI companionship load negatively; Medication, Meditation, Mood tracking, and Gamified self-improvement load positively. That is, communities

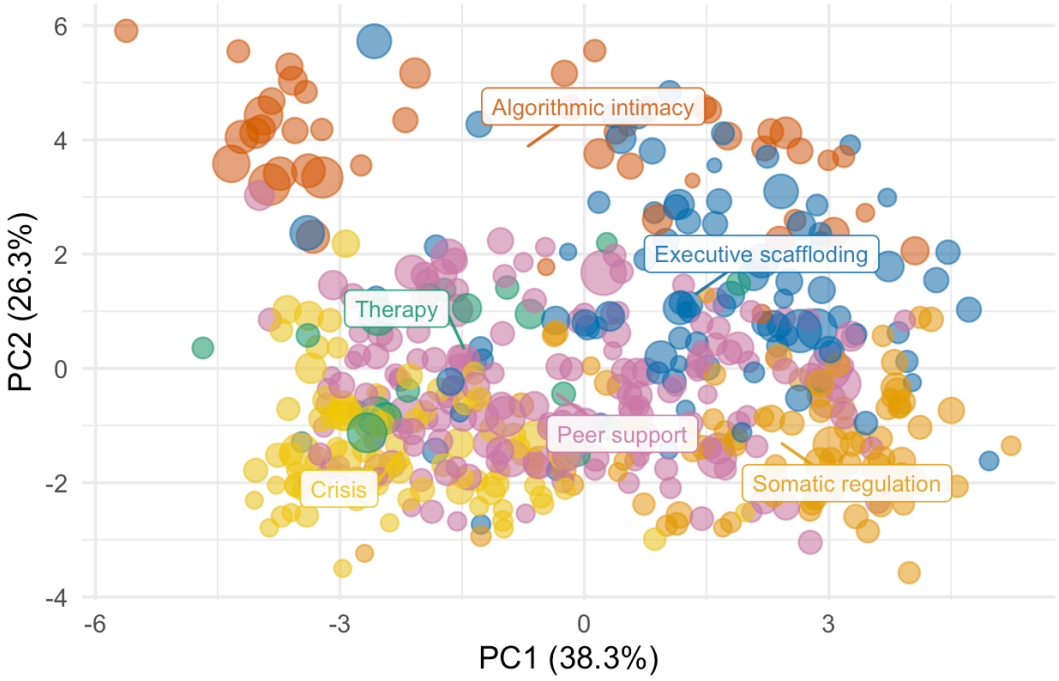


Figure 3: PCA plot (64.6% variance) of subreddits coloured by cluster membership. Point size reflects subreddit corpus size. PC1 separates relational exposure or need (left) from self-governance (right); PC2 separates clinical authority (bottom) from platform-native care (top).

on one end discuss MHT in a context of relational support; communities on the other use it as part of self-regulation or improvement practices.

PC2 (26.3% variance) distinguishes clinical authority from platform-native care. Professional therapy, Medication, and Therapeutic techniques load negatively; AI companion platforms and Gamified self-improvement load positively.

The resulting PCA plot (Figure 3) positions each subreddit as a point, coloured by cluster membership. The six clusters are broadly spatially proximate when characterised by topic mixture alone. But the plot also reveals internal heterogeneity within clusters (subreddits within the same category may be pulled toward different regions) and boundary zones where communities from different clusters intermingle. This variation is analytically productive: it shows where meaning is contested or where discourses blend.

6 Worked Example: What the Method Reveals

Applied to MHT discourse, the pipeline reveals that the same technologies are constituted differently across community contexts. The following analysis, first, makes use of qualitative readings of hundreds of θ-distribution exemplar posts per category, and second, recognises differences within PCA space.

The Peer support cluster occupies a central position, mediating between the others. Its location reflects its general role as an entry point or general space for support, while its dispersion reflects the varied needs of users.

The Therapy cluster, to a lesser extent, also sits centrally, reflecting the way clinical care remains central even in MHT discourse. Its opposition to Algorithmic intimacy along PC2 demonstrates that professional care and AI technology occupy opposite poles.

The Crisis cluster occupies the lower-left: strongly relational and strongly tethered to clinical

authority. Its diagonal opposition to Executive scaffolding marks a fundamental tension in our two dimensions: care as relational and clinically legitimate versus care that is platform-based and self-optimised.

The Somatic regulation cluster occupies the lower-right: high self-regulation combined with strong clinical oversight. Its diagonal opposition to Algorithmic intimacy is, again, instructive: here, bodies are undergoing treatment and pharmacological adjustment under clinical authority; there, affective needs are addressed through synthetic and disembodied relationships.

The Executive scaffolding cluster occupies the upper-right: self-governance combined with platform-based infrastructure.

While the Algorithmic intimacy cluster moves across PC1, there is a noticeable split on the X-axis between AI model and companion focused subreddits (on the left) and a scattering of more general technology subreddits (on the right). The strong presence in the upper-left of these AI subreddits pulls the centroid in that direction. There, MHT is relational but platform-based; care is independent of clinical authority (in opposition to the Therapy cluster).

What we can see here is how substantive discourse shifts between contexts, despite often speaking about the same technologies. For example, a mood tracking app discussed in Somatic regulation functions as a lay clinical instrument, where graphs and logs track various metrics related to their recoveries. The same app in the Executive scaffolding cluster, however, is used to externalise daily regulation (such as doing chores or remembering to drink water) when internal capacities are seen as unreliable. These are not merely different attitudes toward the same object. They reflect fundamentally different models of what care is, where authority resides, and how responsibility is allocated. The method makes visible this variation because it treats topic distributions as discursive signatures rather than thematic labels, and because it operates at the community and individual document level together.

7 Transferability and Limitations

This method is designed to be transferable beyond this specific case study. Its core logic applies wherever researchers wish to study how cultural artefacts are socially constructed across communities.

Several decisions require adaptation for new domains. Keyword design must balance recall (capturing relevant discourse) against precision (excluding false positives); the level of aggregation (subreddit, forum, hashtag community) depends on platform architecture; the number of topics requires domain-specific evaluation; and the manual categorisation of communities and analysis of exemplar posts, while labour-intensive, provides the ground-truth against which computational patterns are validated.

A number of limitations are important to note. Keyword-based sampling excludes relevant discourse that does not explicitly reference target terms. Aggregation at the community level obscures individual trajectories and cross-community participation. Reddit's demographic composition limits generalisability to broader populations. The method captures synchronic structure but not temporal dynamics, such as how discursive regimes emerge, stabilise, or fragment over time. Finally, the interpretive step (reading high-θ documents, labelling topics, and clusters) remains labour-intensive and necessarily shaped by the researcher's analytic framework.

8 Conclusion

This paper has presented a computational method for studying how meaning varies across online communities. By treating topic-proportion vectors as discursive signatures and aggregating them at the community level, the method produces a topology of discourse; a spatial representation in which proximity represents discursive similarity and distance represents divergence. Applied to

MHT discourse on Reddit, it reveals six distinct regimes through which technologies emerge depending on community context. The method integrates individually established techniques (topic modelling, hierarchical clustering, PCA) into a coherent analytical pipeline for studying contextual meaning-making at scale. It is transferable to other domains where cultural artefacts, technologies, or practices carry different meanings across the social spaces in which they are discussed.

References

[1] Ammari, Tawfiq, Schoenebeck, Sarita, and Romero, Daniel M. "Pseudonymous Parents: Comparing Parenting Roles and Identities on the Mommit and Daddit Subreddits". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–13. DOI: 10.1145/3173574.3174063. URL: https://doi.org/10.1145/3173574.3174063.

[2] Anderson, Marti J. "Permutational Multivariate Analysis of Variance (PERMANOVA)". In: *Wiley StatsRef: Statistics Reference Online*, ed. by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. John Wiley & Sons Ltd., 2017. DOI: 10.1002/9781118445112.stat07841. URL: https://doi.org/10.1002/9781118445112.stat07841.

[3] Aranda, Ana Maria, Sele, Kathrin, Etchanchu, Hélène, Guyt, Jelle Y., and Vaara, Eero. "From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling". In: *European Management Review 18*, no. 3 (2021), pp. 197–214. DOI: 10.1111/emre.12452.

[4] Biber, Douglas, Egbert, Jesse, and Keller, Daniel. "Reconceptualizing register in a continuous situational space". In: *Corpus Linguistics and Linguistic Theory 16*, no. 3 (2020), pp. 581–616. DOI: 10.1515/cllt-2018-0086.

[5] Bourdieu, Pierre. *Outline of a Theory of Practice*. Cambridge University Press, 1977. DOI: 10.1017/CBO9780511812507.

[6] De Choudhury, Munmun and De, Sushovan. "Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity". In: *Proceedings of the International AAAI Conference on Web and Social Media 8*, no. 1 (2014), pp. 71–80. DOI: 10.1609/icwsm.v8i1.14526.

[7] Denny, Matthew J. and Spirling, Arthur. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It". In: *Political Analysis 26*, no. 2 (2018), pp. 168–189. DOI: 10.1017/pan.2017.44.

[8] En, Boka, En, Michael, and Griffiths, David. "Gay Stuff and Guy Stuff: The Construction of Sexual Identities in Sidebars on Reddit". In: *Networking Knowledge: Journal of the MeCCSA Postgraduate Network 6*, no. 1 (2013), pp. –. DOI: 10.31165/nk.2013.61.293. URL: https://doi.org/10.31165/nk.2013.61.293.

[9] Geertz, Clifford. *The Interpretation of Cultures*. Basic Books, 1973.

[10] Goodings, Lewis, Ellis, Darren, and Tucker, Ian. *Understanding Mental Health Apps: An Applied Psychosocial Perspective*. Springer Nature, 2024.

[11] Miller, Vincent, Hill, Mark J., and Moreira, Tiago. "Artificial Intelligence and Unconditional Love: The Rise of Generative AI as an Alternative Form of Mental Health Support". In: *Communication and Change* (2026). Accepted manuscript, accepted 28 June 2026. Forthcoming.

[12] Mimno, David, Wallach, Hanna, Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[13] Moreira, Tiago, Miller, Vince, and Hill, Mark J. “Two kinds of mirror?: recrafting mediated selves in apps for mental wellbeing”. In: *SSM - Qualitative Research in Health 10* (2026), p. 100818. ISSN: 2667-3215. DOI: https://doi.org/10.1016/j.ssmqr.2026.100818. URL: https://www.sciencedirect.com/science/article/pii/S2667321526001204.

[14] Nelson, Laura K. “Computational Grounded Theory: A Methodological Framework”. In: *Sociological Methods & Research 49*, no. 1 (2020), pp. 3–42. DOI: 10.1177/0049124117729703.

[15] Park, Albert and Conway, Mike. “Harnessing Reddit to Understand the Written-Communication Challenges Experienced by Individuals With Mental Health Disorders: Analysis of Texts From Mental Health Communities”. In: *Journal of Medical Internet Research 20*, no. 4 (2018), e121. DOI: 10.2196/jmir.8219. URL: https://doi.org/10.2196/jmir.8219.

[16] Park, Albert, Conway, Mike, and Chen, Annie T. “Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach”. In: *Computers in Human Behavior 78* (2018), pp. 98–112. DOI: 10.1016/j.chb.2017.09.001.

[17] Pilkington, Pamela D. and Rominov, Holly. “Fathers’ Worries During Pregnancy: A Qualitative Content Analysis of Reddit”. In: *The Journal of Perinatal Education 26*, no. 4 (2017), pp. 208–218. DOI: 10.1891/1058-1243.26.4.208. URL: https://doi.org/10.1891/1058-1243.26.4.208.

[18] Roberts, Margaret E., Stewart, Brandon M., and Tingley, Dustin. “stm: An R Package for Structural Topic Models”. In: *Journal of Statistical Software 91* (2019), pp. 1–40. DOI: 10.18637/jss.v091.i02.

[19] Squirrell, Tim. “Platform dialectics: The relationships between volunteer moderators and end users on reddit”. In: *New Media & Society 21*, no. 9 (2019), pp. 1910–1927. DOI: 10.1177/1461444819834317.

[20] Suchman, Lucy. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511808418.

[21] Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. “Evaluation methods for topic models”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 1105–1112. DOI: 10.1145/1553374.1553515.

[22] Ward, Joe H. “Hierarchical Grouping to Optimize an Objective Function”. In: *Journal of the American Statistical Association 58*, no. 301 (1963), pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

[23] Williams, Raymond. *Keywords: A Vocabulary of Culture and Society*. Routledge, 2011.

[24] Yang, Kailai, Zhang, Tianlin, Kuang, Ziyan, Xie, Qianqian, Huang, Jimin, and Ananiadou, Sophia. “MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models”. In: *Proceedings of the ACM Web Conference 2024*. Association for Computing Machinery, 2024, pp. 4489–4500. DOI: 10.1145/3589334.3648137. URL: https://doi.org/10.1145/3589334.3648137.