

Detecting sub-populations in online health communities: A mixed-methods exploration of breastfeeding messages in BabyCenter Birth Clubs

Calla Beauregard,^{1,2,*} Parisa Suchdev,^{1,2,3} Ashley M. A. Fehr,^{1,2} Isabelle T. Smith,¹ Tabia Tanzin Prama,^{1,2} Julia Witte Zimmerman,^{1,2,3} Carter Ward,³ Juniper Lovato,^{1,4,3} Christopher M. Danforth,^{1,2,5} and Peter Sheridan Dodds^{1,2,4,6,†}

¹*Vermont Complex Systems Institute, University of Vermont, Burlington, VT 05405, USA.*

²*Computational Story Lab, University of Vermont, Burlington, VT 05405, USA.*

³*Computational Ethics Lab, University of Vermont, Burlington, VT 05405, USA.*

⁴*Department of Computer Science, University of Vermont, Burlington, VT 05405, USA.*

⁵*Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA.*

⁶*Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA.*

(Dated: October 30, 2025)

Parental stress is a nationwide health crisis according to the U.S. Surgeon General’s 2024 advisory. To allay stress, expecting parents seek advice and share experiences in a variety of venues, from in-person birth education classes and parenting groups to virtual communities, for example, BabyCenter, a moderated online forum community with over 4 million members in the United States alone. In this study, we aim to understand how parents talk about pregnancy, birth, and parenting by analyzing 5.43M posts and comments from the April 2017–January 2024 cohort of 331,843 BabyCenter “birth club” users (that is, users who participate in due date forums or “birth clubs” based on their babies’ due dates). Using BERTopic to locate breastfeeding threads and LDA to summarize themes, we compare documents in breastfeeding threads to all other birth-club content. Analyzing time series of word rank, we find that posts and comments containing anxiety-related terms increased steadily from April 2017 to January 2024. We used an ensemble of topic models to identify dominant breastfeeding topics within birth clubs, and then explored trends among all user content versus those who posted in threads related to breastfeeding topics. We conducted Latent Dirichlet Allocation (LDA) topic modeling to identify the most common topics in the full population, as well as within the subset breastfeeding population. We find that the topic of sleep dominates in content generated by the breastfeeding population, as well anxiety-related and work/daycare topics that are not predominant in the full BabyCenter birth club dataset.

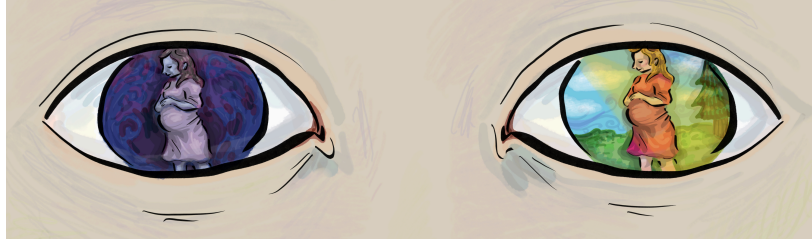


FIG. 1. Figure representing the complex milieu of emotions surrounding fertility, pregnancy, and birth, even within a single person’s experience.

I. Introduction

The Centers for Disease Control (CDC) in the United States has reported that American women are largely dissatisfied with the quality of their maternity care [1]. Based on the largest survey to date, 1 in 5 women report mistreatment during their care. Racial disparity compounds reports of mistreatment, with 1 in 3 Black, Hispanic, and multiracial women reporting mistreatment.

Furthermore, 45% of women report holding back when asking questions of their maternity care providers [1]. Given that demographics (specifically racial identity) can impact endorsement of medical symptoms by providers, at least under some circumstances [2],[3] it is plausible that there is interaction between identity and the quality of an individual’s patient-provider interaction.

In addition to poor experience with maternity care, maternal mortality rates are higher in the United States than in comparable nations [4], and particularly high among Black, non-Hispanic women as compared to White women [5].

* calla.beauregard@uvm.edu

† peter.dodds@uvm.edu

In order to understand questions of public health, researchers often seek out the ‘highest quality evidence’ in the form of randomized controlled trials, but such studies are expensive [6] and have historically excluded women [7]. Public health research has increasingly focused on big data sources and employed epidemiological techniques to better understand populations [8]. Practicing “digital epidemiology” (epidemiology on non-traditional data sources, see Salathé’s 2018 review [9]) means using publicly available data to identify trends invisible in more traditional formats. Social media data, which often chronicles day-to-day mood and activity, has particular promise for mental health research [10]. However, there is potential for significant bias in individuals’ online self-portrayals, as well as ethical questions related to research use of data shared by users who may not have intended it for such purposes [11].

Considering the recent turmoil surrounding women’s health funding in the United States and funding for foreign aid programs [12, 13], it is vital that public health researchers consider other avenues to understand populations, especially those that historically underserved. Previous work has explored language differences in parenting subreddits between mothers and fathers using topic modeling [14] as well as explored discourse on Mumsnet, a United Kingdom based parenting forum, using an anthropological approach [15]. In this work, we employ mixed methods and ensemble topic modeling to identify trends in language associated with users who post about breastfeeding on BabyCenter, a moderated online forum community with over 4 million members in the United States alone. We specifically examine breastfeeding due to its important relationship with maternal mental health. Anxiety disorders are more common in postpartum individuals than in the general population, treatment rates are low for postpartum anxiety, and the experience of postpartum anxiety can be associated with early breastfeeding cessation [16]. Accordingly, recent research demonstrates a significant relationship between the duration of breastfeeding and subjective norms [17], guilt and poor maternal mental health [18], and guilt and the shorter duration of exclusive breastfeeding [19]. A cross-sectional online survey ($N = 470$) of mothers 6- and 12-month postpartum suggests that “postpartum anxiety may be an underlying mechanism which reduces exclusive breastfeeding duration and negatively affects maternal perceptions of infant sleep quality” [17]. A lack of studies on postpartum mental health [16], and evidence that anxiety disorders are more prevalent and burdensome in women than in men further motivates the study of the impact and manifestation of anxiety with respect to other aspects of women’s health [2]. [20]

First, we conduct topic modeling using the **BERTopic** Python package [21] to understand the nature of posts and comments in the April 2019 birth club, the largest such club on BabyCenter. We identify the most promi-

nent topics related to breastfeeding across all clubs. We then use Latent Dirichlet Allocation (LDA) to examine trends in language use across both the entire dataset (April 2017–January 2024) and the subset of users who post in breastfeeding topics. We conclude with an analysis of topics inferred from word rank time series and allotaxonomy [22].

II. Description of datasets

The present study uses posts and comments from all US BabyCenter monthly birth clubs spanning April 2017–January 2024 ($n = 5,433,338$ posts and comments by $n = 331,843$ unique users). Each birth club represents a cohort of US-based BabyCenter users with expected due dates in that club’s given month and year. Users post the majority of their content in the months leading up to the delivery date (their club’s namesake), with activity sometimes continuing but generally tapering off after delivery.

The dataset contains a unique username identifier and the calendar date for each post or comment. For clarity, we examine all posts and comments and will refer to them as “documents” for the remainder of this paper. Figure 2 shows the distribution of word counts across all documents in the dataset.

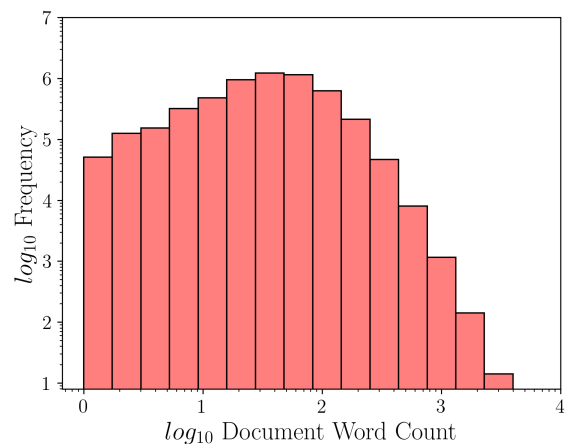


FIG. 2. **Histogram of Words of All Documents using kernel density estimation (KDE) for smoothing.** The most common documents (posts and comments), appearing approximately 1 million times in the dataset, are roughly 30-40 words each, displayed by the bin height.

To validate longitudinal temporal trends in user posts and comments in birth clubs, we first use allotaxonomy (the study of rank turbulence, see Dodds et al. 2023 [22]). We combine data from every birth club based on trimester thresholds using the first day of the due

date month, and partition the data by three-month increments. We create groups from this data transformation to represent all posts in the first trimester and all posts in the fourth trimester. We then compare these groups using the Allotaxonomer. In Figure 3, we observe that words like “ultrasound” appear more frequently by rank in the first trimester but pronouns like “her”, “she” and “him” as well as the word “baby” occur more frequently in the fourth trimester. Accordingly, we consider birth clubs as “cohorts” for this study, and assume that each club contains longitudinal discussion of pregnancy and birthing.

III. Methods

A. Identifying sub-populations with ensemble topic modeling

We use an ensemble of topic modeling, starting with BERTopic, to find longitudinal themes in posts and comments across birth club cohorts, specifically in populations discussing breastfeeding. According to its developers, “by default, the main steps for topic modeling with BERTopic are sentence-transformers, UMAP, HDBSCAN, and c-TF-IDF run in sequence” [21]. Essentially, sentence transformers embed textual data in matrices, and UMAP (or Uniform Manifold Approximation) performs dimensional reduction of the high-dimensional embedding matrices [23]. HDBSCAN (or Hierarchical Density-Based Spatial Clustering of Applications with Noise) determines high-density clusters embedding [24], and the c-TF-IDF (Class-based Term Frequency-Inverse Document Frequency) returns terms and topics based on relative presence in documents [21]. We retain the order of these steps for this analysis and use a random seed of 42 in order to replicate the stochastic aspects of topic modeling. We then identify the top breastfeeding topic in each birth club through keyword search of topic names and representative documents in each topic. We are highly inclusive in our selection and ensure that we include terms like “ebf” (i.e., exclusive breastfeeding) and “pumping” as alternate breastfeeding terms. We aim to select the largest breastfeeding topic so long as it was topically related to breastfeeding experience. For example, some breastfeeding topics are primarily about “nursing bras”. In this case, we select the next largest, topically related breastfeeding topic, e.g., “infant feeding”.

We first analyze each birth club using topic analysis and then qualitatively code documents within topics to further sample the population using three human coders. Employing this labeled dataset, we test a variety of common machine learning classifiers using a random 80/20 train/test split of the annotated data. We found that

the classifiers did not perform well, even when taking the union of agreement between the three best classifiers and conducting human annotation. Essentially, the classifiers disagreed on different posts and comments with little to no interpretability, despite reasonably high accuracy scores. For completeness, we report our qualitative coding scheme and accuracy scores in the Appendix (see Table A8). Based on this lack of interpretability, we employed a simpler approach and subset the population by all users who posted in the largest breastfeeding topic in each birth club. We include all documents in all topics in the particular birth club for each user who posted in the breastfeeding topic.

B. Analysis of topics and breastfeeding sub-population

We adopt an ensemble approach to topic modeling, combining the strengths of multiple algorithms to improve classification accuracy and interpretability. This method follows prior work demonstrating the effectiveness of topic modeling ensembles [25]. For the initial document labeling, we employ BERTopic, which leverages contextual embeddings and clustering techniques. This method is particularly suitable for natural language processing (NLP) tasks requiring a nuanced understanding of document semantics.

We then apply Latent Dirichlet Allocation (LDA), a widely used generative probabilistic model that discovers latent topics in discrete data sets such as text corpora [26, 27]. The primary goal of topic modeling is to uncover the main themes present in unstructured textual data [28]. LDA groups words based on semantic similarity [29]. Each word within a topic is associated with a conditional probability, indicating its relevance to the respective topic cluster. The resulting word clusters represent distinct themes within the corpus. A limitation of the LDA model is that it does not inherently assign descriptive labels to the topics it generates [30]. To address this, we manually label each topic by interpreting its top keywords. We perform a sweep across the best number of topics (k) and two distributional parameters (α and η) to determine the hyperparameters that statistically yield the best goodness of fit via perplexity [27, 31, 32] (refer to Appendix Table A9 for all perplexity score results). In our implementation, we extract the top 30 keywords across 20 distinct topics based on the lowest perplexity score across the entire dataset. To ensure accurate labeling, domain experts review and interpret the keywords associated with each topic, which remains the most effective method for interpreting unlabeled topic model outputs. Subsequently, we construct topic-document matrices, where each entry represents the probability of a document belonging to a specific topic. These matrices

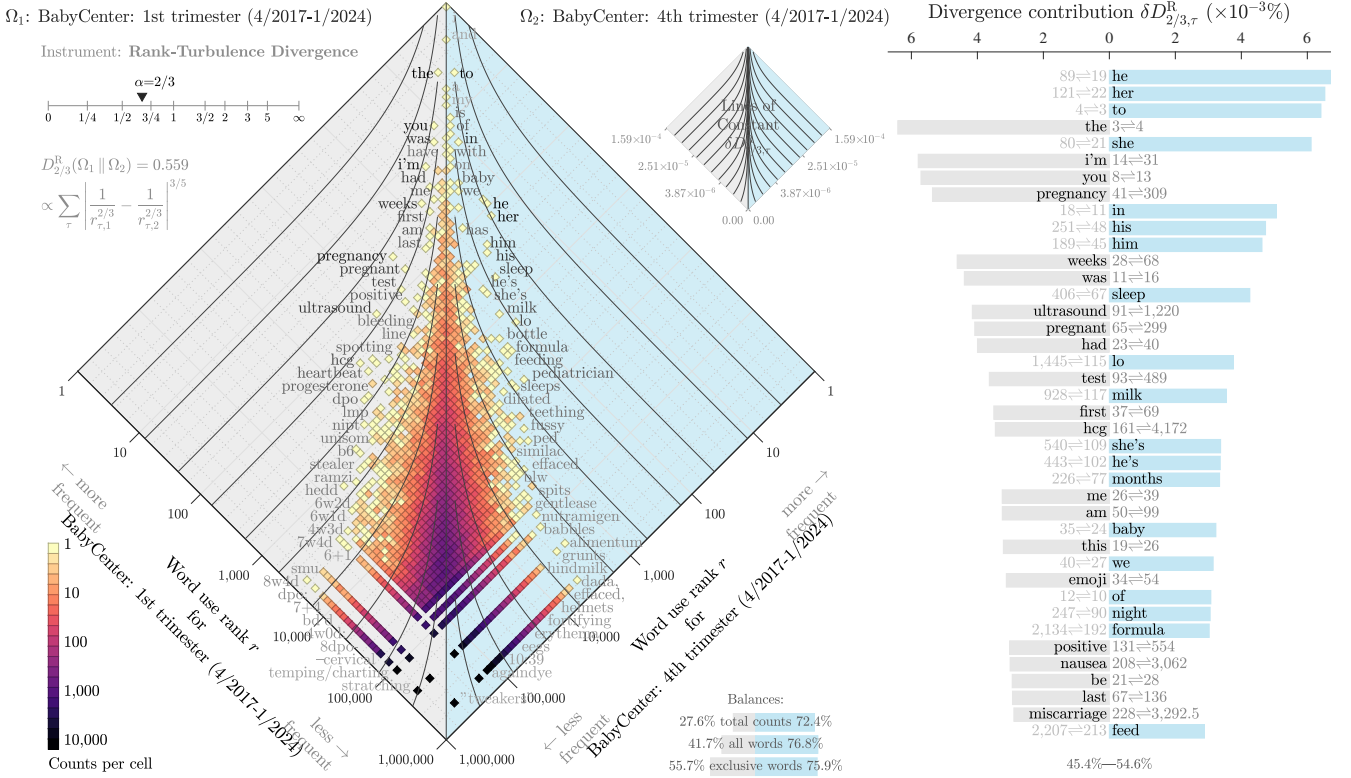


FIG. 3. **Allotaxonomic Representation of First and Fourth Trimesters, April 2017–January 2024.** In the Allotaxonograph, the first trimester appears on the left-hand panel, and the fourth trimester appears on the right-hand panel. Reading from top to bottom, the highest-ranked words in each corpus appear sequentially in descending order.

ces facilitate the identification of representative keywords for each topic and their association with individual documents. We use these same parameter selections for both the entire dataset, and the subset for consistency.

IV. Results

In Figure 4, both the largest and smallest topics by birth club are general breastfeeding content. We manually verify the largest topics related to breastfeeding by inspecting representative documents to limit the inclusion of tangentially related documents. We note that there is still considerable variation across topics in birth clubs due to the stochasticity inherent in UMAP clustering and other features of topic modeling. Unfortunately, this limitation is only verifiable upon manually inspecting every document. In Figure 5, we further observe that the distribution of word frequencies across the subset of breastfeeding documents has a similar shape to the distribution of word frequencies by document in the full corpus. This suggests that our subset of breastfeeding posts and comments does not deviate substantially in content length from the full dataset.

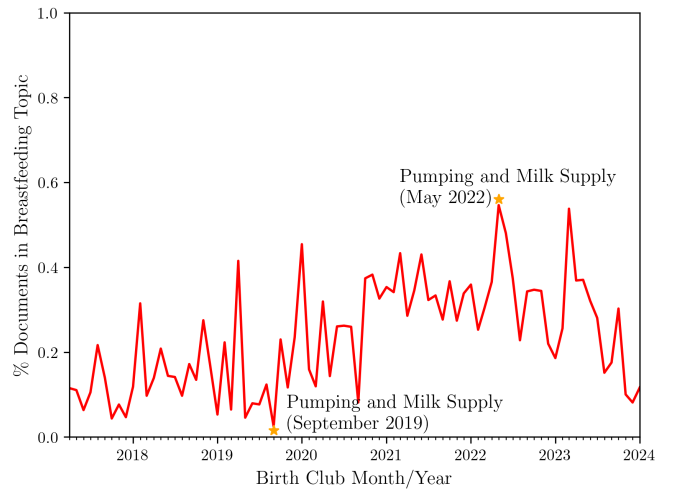


FIG. 4. **Time Series of Percentage of Total Documents for Dominant Breastfeeding Topics, Birth Clubs April 2017–January 2024.** The proportion of breastfeeding topics in monthly birth clubs vacillates over time. The birth clubs with the largest (May 2022) and smallest (September 2019) breastfeeding topics by percent share are marked with a star.

A. Trends in word use in full population and breastfeeding sub-population

From April 2017–June 2023, terms related to anxiety consistently increase across orders of magnitude, demonstrated in Figure 6. We exclude July 2023–January 2024 birth clubs in the plot to ameliorate any truncated birth clubs that have limited follow-up, as the dataset ends in January 2024. We derived these words using synonyms for anxiety from the Diagnostic Statistical Manual 5th edition (DSM-V) [33].

B. Trends in topics in full population and breastfeeding sub-population

We perform LDA on the full dataset and compare its topics to those of the sub-population. For the full dataset, posts and comments in the “Work/Family” topic dominate proportionally across April 2017–June 2023 birth clubs. A “Baby Sleep” topic ranks 9th in the full dataset, while it ranks 4th in the breastfeeding subset. Additionally, while “anxiety” does not emerge as a topic in the full dataset, the “anxiety”-related topic in the sub-population follows the same increasing importance as word rank. We also notice a “Family/work/leave” topic in the breastfeeding group that is not present in the total dataset group. Considering the importance of workplace accommodation in breastfeeding and pumping after returning to work, users participating in breastfeeding threads discuss work/leave more than all users.

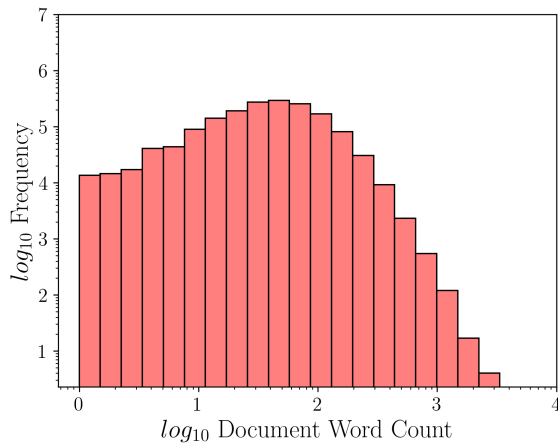


FIG. 5. **Histogram of Words of Breastfeeding Documents using Kernel Density Estimation (KDE) for Smoothing.** The most common documents (posts and comments), appearing approximately 300,000 times in the breastfeeding subset, are roughly 30-40 words each, displayed by the bin height. This shape is very similar to the distribution for the total birth club dataset.

V. Discussion

Our mixed topic modeling scheme identifies a subset of documents related to breastfeeding, generating a novel dataset from which to conduct more traditional public health research. A non-exhaustive review of these documents reveals a breadth of sentiment and experience with breastfeeding that is reflected across other forms of research [34–36]. Recent research has associated non-exclusive breastfeeding (that is, breastfeeding supplemented with formula) to higher levels of anxiety and depression in a study of 229 women who were followed pre-birth to 3–6 months postpartum [35]. Topic headings generated by topic modeling (either LDA or BERTopic) can obfuscate important information contained in posts/comments, making it difficult to correctly classify nuanced text reflecting specific shame or embarrassment related to breastfeeding [37]. However, the absence of *any* “anxiety” or “leave” topics in the full dataset topics strongly suggests that users who post in the breastfeeding topics comment more readily on anxiety, as they may experience more anxiety compared to the entire population posting in birth clubs. Furthermore, the specific inclusion of “leave” in the breastfeeding documents suggests the importance of workplace accommodation in breastfeeding and pumping after returning to work.

Based on the increasing understanding of the interplay between mental health and breastfeeding, the relative prominence of sleep-related topics in the breastfeeding sub-population versus the general population is also worth exploring further. The positive impact of adequate sleep on mental health has been well-established [38], which may explain why sleep discussions are prevalent in a group that self-reports many anxious feelings. However, without establishing causality, it is also possible that poor sleep may have contributed to breastfeeding struggles, and sleep improvement interventions may be valuable tools to increase the likelihood of continued breastfeeding. Likewise, the prominence of leave-related discussions in the breastfeeding struggles sub-population may suggest interventions that improve workplace accommodation for breastfeeding may be worth prioritizing as a target for improvement [39–41].

More globally, our results demonstrate a method for examining discussion in online communities for use in public health research. Using topic modeling, distant reading of online communities with NLP can supplement and enhance survey data. Additionally, since this data arises from a self-selected internet community, it is possible that it is more self-reflective and data-rich, as indicated by studies on internet surveys [42]. Furthermore, traditional surveys usually occur at a single point in time, in a specific setting (e.g., a screening form in a doctor’s office), while our research considers a longer

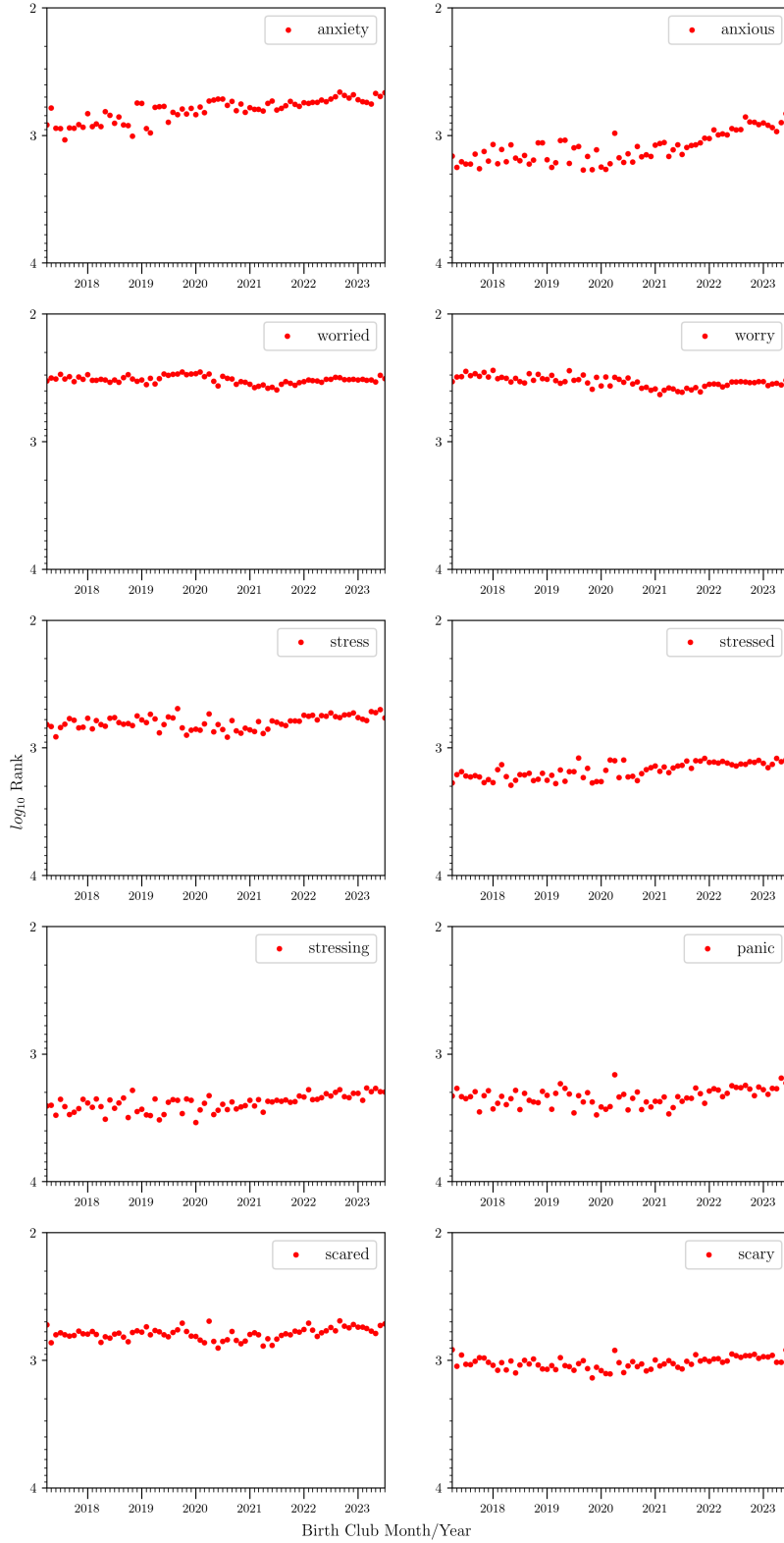


FIG. 6. Rank Time Series of Diagnostic Statistical Manual (DSM) terms for “anxiety” and similar terms, BabyCenter birth clubs (April 2017–June 2023). Anxiety terms increase over time across birth clubs, with the terms “anxiety” and “anxious” showing the greatest increases over magnitudes.)



FIG. 7. **Normalized Topic Proportions by BabyCenter Birth Club.** These panels represent the topic proportions ranked by size of the top 20 topics in the full birth club dataset (left) and of all posts and comments by users who posted in dominant breastfeeding topics (right). Shared topics across both datasets have the same color. While “Family/Work” appears across both groups, “anxiety” only appears as a dominant topic for the breastfeeding subpopulation.

context window. Our approach considered longitudinal information across pregnancy and the postpartum period, thus providing a larger context window regarding birthing and parenting experiences, with our results suggesting important concerns for expecting and new parents that one-time measures may miss. Expanding the context window longitudinally can further seed topics of concern to ask about in surveys or doctor office intake forms. Additionally, surveys and controlled studies often focus on subpopulations of motivated participants who are convenience-sampled at the study’s location, particularly when studies have in-person or clinical components, which can introduce a source of bias. The methods we follow to curate a sub-population allow opt-in of many concerned parties that may otherwise not have access to these studies.

However, it is important to underscore some of the limitations of this study. First, using BERTopic depends on a pipeline of distinct algorithms, which introduces reproducibility issues for future researchers. For instance, as evidenced by the alternative topic models provided in the Appendix, UMAP yields different clusters in each instance of topic analysis and how we decide to represent them can lead to varied interpretations. Secondly, self-disclosed information collected in a non-research environment is not subject to the controls of well-designed research studies and should be closely examined for bias and missing data.

VI. Concluding remarks

Considering the limitations of traditional surveys (measurement, coverage, and response bias) [43, 44], as well as reluctance by birthing people to relate issues to their providers [1], exploring online communities for risk and protective factors for various conditions is an effective time- and cost-saving venture. Our work makes a first step in identifying an important sub-population of breastfeeding people, employing an ensemble of NLP and topic modeling techniques to more expansively study concerns in this sub-population.

VII. Ethics Statement

This study analyzed publicly available data from the US BabyCenter website, a large online health and parenting community. In compliance with our institutional guideline the project was reviewed using the university’s human subjects determination tool and was determined not to constitute human subjects research. Therefore, it did not require formal Institutional Review Board (IRB) review or approval. All analyses complied with BabyCenter’s Terms of Service and respect for user privacy. Posts were accessed in aggregate, and no attempts were made to contact users or link usernames across platforms. To further preserve anonymity, no quoted text is revealed to remove potentially identifying details, and no direct user handles are reported.

VIII. Data Availability

The data used in this study were derived from publicly accessible discussion forums hosted on BabyCenter (www.babycenter.com) and were collected in compliance with the site’s Terms of Service. Because the raw text of posts may contain potentially identifying or sensitive personal information, we do not make raw text data publicly available. Researchers interested in reproducing analytic steps may request derived, de-identified metadata (e.g., topic assignments, document counts, or aggregated term frequencies) and analysis code from the corresponding author upon reasonable request.

Acknowledgments

We are grateful for conversations with Bradford Demarest on Topic Analysis. Additionally, we thank Carter Ward, Sarah Nowak, and Jay Lobell for their management of the data from which this was built.

Our work was supported in part by MassMutual, National Science Foundation Award #242829 (Science of Online Corpora, Knowledge, and Stories), and a philanthropic gift from an anonymous source. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the aforementioned supporters.

[1] C. for Disease Control and Prevention, [Mistreatment during maternity care](#) (2023).

[2] C. P. McLean, A. Asnaani, B. T. Litz, and S. G. Hof-

- mann, Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness, *Journal of Psychiatric Research* **45**, 1027 (2011).
- [3] McLean *et al.* [2] discusses psychopathology specifically, and since mental health – like motherhood – is an intimate, value-laden, culturally-mediated, even stigmatized, health topic, it seems a likely comparison in this regard.
 - [4] R. Tikkanen, M. Z. Gunja, M. FitzGerald, and L. C. Zephyrin, *Maternal Mortality and Maternity Care in the United States Compared to 10 Other Developed Countries* (2020).
 - [5] D. L. Hoyert, Maternal Mortality Rates in the United States 2022, *NCHS Health E-Stats* (2024).
 - [6] B. Speich, B. von Niederhäusern, N. Schur, L. G. Hemkens, T. Fürst, N. Bhatnagar, R. Alturki, A. Agarwal, B. Kasenda, C. Pauli-Magnus, M. Schwenkglenks, and M. Briel, Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data, *Journal of Clinical Epidemiology* **96**, 1 (2018).
 - [7] K. A. Liu and N. A. D. Mager, Women’s involvement in clinical trials: historical perspective and future implications, *Pharmacy Practice* **14**, 708 (2016).
 - [8] S. Dolley, Big Data’s Role in Precision Public Health, *Frontiers in Public Health* **6**, 68 (2018).
 - [9] M. Salathé, Digital epidemiology: what is it, and where is it going?, *Life Sciences, Society and Policy* **14**, 1 (2018).
 - [10] A. M. Stupinski, T. Alshaabi, M. V. Arnold, J. L. Adams, J. R. Minot, M. Price, P. S. Dodds, and C. M. Danforth, Quantifying Changes in the Language Used Around Mental Health on Twitter Over 10 Years: Observational Study, *JMIR Mental Health* **9**, e33685 (2022), company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.
 - [11] M. Tušl, A. Thelen, K. Marcus, A. Peters, E. Shalaeva, B. Scheckel, M. Sykora, S. Elayan, J. A. Naslund, K. Shankardass, S. J. Mooney, M. Fadda, and O. Gruebner, Opportunities and challenges of using social media big data to assess mental health consequences of the COVID-19 crisis and future major events, *Discover Mental Health* **2**, 14 (2022).
 - [12] G. Grossi, *HHS Cuts Funding for NIH-Based Women’s Health Initiative Threatening Decades-Long Study* (2025).
 - [13] A. N. Kallen, S. Whirlledge, K. N. Goldman, and J. Johnson, Undermining Women’s Health Research — Gambling with the Public’s Health, *New England Journal of Medicine* **392**, 2185 (2025), publisher: Massachusetts Medical Society eprint: <https://www.nejm.org/doi/pdf/10.1056/NEJMp2503576>.
 - [14] M. Sepahpour-Fard and M. Quayle, How do mothers and fathers talk about parenting to different audiences?: Stereotypes and audience effects: An analysis of r/daddit, r/mommit, and r/parenting using topic modelling, in *Proceedings of the ACM Web Conference 2022 (WWW ’22)* (ACM, Virtual Event, Lyon, France, 2022) p. 11, april 25–29, 2022.
 - [15] A. Locke, Book review: Language, gender and parenthood online: Negotiating motherhood in mumsnet talk by jai mackenzie, *Feminism & Psychology* **32**, 125 (2021).
 - [16] E. Ali, Women’s experiences with postpartum anxiety disorders: A narrative literature review, *International Journal of Women’s Health* **10**, 237 (2018).
 - [17] S. M. Davies, B. F. Todd-Leonida, V. M. Fallon, and S. A. Silverio, Exclusive Breastfeeding Duration and Perceptions of Infant Sleep: The Mediating Role of Postpartum Anxiety, *International Journal of Environmental Research and Public Health* **19**, 4494 (2022).
 - [18] L. Jackson, L. De Pascalis, J. Harrold, and V. Fallon, Guilt, shame, and postpartum infant feeding outcomes: A systematic review, *Maternal & Child Nutrition* **17**, e13141 (2021).
 - [19] P. S. Russell, M. D. Birtel, D. M. Smith, K. Hart, and R. Newman, Infant feeding and internalized stigma: The role of guilt and shame, *Journal of Applied Social Psychology* **51**, 906 (2021).
 - [20] About one in three women will meet the criteria for anxiety disorders during their lifetime, versus roughly one in five men [2].
 - [21] M. Grootendorst, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (2022), arXiv:2203.05794 [cs].
 - [22] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth, *Allotaxonomy and rank-turbulence divergence: A universal instrument for comparing complex systems* (2023), arXiv:2002.09770 [physics].
 - [23] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv **1802.03426v3** (2020).
 - [24] L. McInnes, J. Healy, and S. Astels, hdbscan: Hierarchical density based clustering, *The Journal of Open Source Software* **2**, 205 (2017).
 - [25] L. George and P. Sumathy, An integrated clustering and BERT framework for improved topic modeling, *International Journal of Information Technology* **15**, 2187 (2023).
 - [26] P. Kherwa and P. Bansal, Topic modeling: A comprehensive review, *EAI Endorsed Transactions on Scalable Information Systems* **7**, 1 (2020).
 - [27] R. Egger and J. Yu, Topic modeling comparison between lda, top2vec, and bertopic to demystify twitter posts, *Frontiers in Sociology* **7**, 1 (2022).
 - [28] N. Rogers and L. Longo, A comparison on the classification of short-text documents using latent dirichlet allocation and formal concept analysis, in *CEUR Workshop Proceedings*, Vol. 2086 (2017) pp. 50–62.
 - [29] D. W. Alkhafaji and S. A. Al-Rashid, Topic modeling for clustering arabic documents, in *Proceedings of the 2nd Information Technology to Enhance E-Learning and Other Applications Conference (IT-ELA 2021)* (2021) pp. 76–81.
 - [30] M. H. Rahman, T. T. Prama, and M. M. Anwar, Modeling topic specific credibility in twitter based on structural and attribute properties, in *International Conference on Health Information Science* (2020).
 - [31] L. Benites-Lazaro, L. Giatti, and A. Giarolla, Topic modeling method for analyzing social actor discourses on climate change, energy and food security, *Energy Research & Social Science* **45**, 318 (2018).
 - [32] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *Latent dirichlet allocation | The Journal of Machine Learning Research* (2003).
 - [33] American Psychiatric Association, *Diagnostic and Sta-*

- tistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)* (American Psychiatric Association, Washington, DC, 2022) text revision of DSM-5, includes updated diagnostic criteria, ICD-10-CM codes, and new disorders such as prolonged grief disorder.
- [34] E. Ystrom, Breastfeeding cessation and symptoms of anxiety and depression: a longitudinal cohort study, [BMC Pregnancy and Childbirth](#) **12**, 36 (2012).
 - [35] S. Coe, M. I. García, A. Mira, and V. Valdés, The Role of Perinatal Anxiety and Depression in Breastfeeding Practices, [Breastfeeding Medicine](#) **15**, 495 (2020), publisher: Mary Ann Liebert, Inc., publishers.
 - [36] E. M. Nagel, M. A. Howland, C. Pando, J. Stang, S. M. Mason, D. A. Fields, and E. W. Demerath, Maternal psychological distress and lactation and breastfeeding outcomes: A narrative review, [Clinical therapeutics](#) **44**, 215 (2022).
 - [37] G. Thomson, K. Ebisch-Burton, and R. Flacking, Shame if you do – shame if you don’t: women’s experiences of infant feeding, [Maternal & Child Nutrition](#) **11**, 33 (2014).
 - [38] A. J. Scott, T. L. Webb, M. Martyn-St James, G. Rowse, and S. Weich, Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials, [Sleep Medicine Reviews](#) **60**, 101556 (2021).
 - [39] Y. Bai and S. M. Wunderlich, Lactation Accommodation in the Workplace and Duration of Exclusive Breastfeeding, [Journal of Midwifery & Women’s Health](#) **58**, 690 (2013), eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmwh.12072>.
 - [40] A. Magner and C. A. Phillipi, Using a Wellness Program to Promote a Culture of Breastfeeding in the Workplace: Oregon Health & Science University’s Experience, [Journal of Human Lactation](#) **31**, 40 (2015), publisher: SAGE Publications Inc STM.
 - [41] S.-Y. Tsai, Impact of a breastfeeding-friendly workplace on an employed mother’s intention to continue breastfeeding after returning to work, [Breastfeeding Medicine: The Official Journal of the Academy of Breastfeeding Medicine](#) **8**, 210 (2013).
 - [42] R. C. Hanna, B. Weinberg, R. P. Dant, and P. D. Berger, Do internet-based surveys increase personal self-disclosure?, [Journal of Database Marketing & Customer Strategy Management](#) **12**, 342 (2005).
 - [43] C. Andrade, The Limitations of Online Surveys, [Indian Journal of Psychological Medicine](#) **42**, 575 (2020).
 - [44] M. Coughlan, P. Cronin, and F. Ryan, Survey research: Process and limitations, [International Journal of Therapy and Rehabilitation](#) **16**, 9 (2009), publisher: Mark Allen Group.
 - [45] H. Szugye, A. Murra, and S. K. Lam, A new policy update on breastfeeding: What all clinicians need to know, [Cleveland Clinic Journal of Medicine](#) **90**, 469 (2023), publisher: Cleveland Clinic Journal of Medicine Section: Guidelines to Practice.

Appendix

Using **BERTopic** utilizes Uniform Manifold Approximation and Projection (UMAP) when clustering topics which employs stochastic gradient descent. Thus, different clusters emerge due to randomness introduced in this step. For clarity, we ran **BERTopic** and use that same mapping throughout our analysis of clusters. However, it is important to note how these clusters can vary based on this step, and a comparison clustering is provided at the end of this section.

Using the standard **BERTopic** steps, we report the top 10 topics that emerged in A7. We disregard the stop word containing topic as is standard using **BERTopic**.

Topic	Count	Name
1	4,226	1_measuring_percentile_oz_lbs
2	2,726	2_induction_induced_induce_39
3	2,604	3_spotting_bleeding_brown_clots
4	2,260	4_movement_kicks_movements_flutter
5	1,798	5_sleep_exhausted_nap_tired
6	1,620	6_dog_cat_dogs_cats
7	1,463	7_gained_gain_weight_pounds
8	1,375	8_section_recovery_vaginal_csection
9	1,330	9_bump_bloat_bloated_showing
10	1,304	10_netflix_book_books_series

TABLE A1. **BERTopic** identified topics in April 2019 birth club.

Within the top 10 identified topics, some deal with early- and mid- pregnancy such as early bump/bloating appearance (topic 9) and kicking/movement (topic 4). Other dominant topics (2, 8) deal with labor itself. A few individuals topics are vague and could refer to before and after pregnancy like measuring and percentiles (topic 1), spotting (3), sleep and tiredness (topic 5), and weight gain (topic 7) which occur during pregnancy and post-birth. Topics that do not occur in a logical grouping are topics about cats and dogs (topic 6) with representative documents speaking to clinginess of pets during pregnancy and Netflix/books (topic 10).

These topics collectively represent the full breadth of concerns across pregnancy and post-birth. Interestingly, breastfeeding does not appear in the broad overview of topics, which may be due to the predominance of other discussions. However, since the Allotaxonomer demonstrated that breastfeeding terms experienced a large shift pre- and post-birth, we then explored breastfeeding specific topics in the data set using the same pipeline previously described. The breastfeeding topics are varied but include conversation of formula feeding versus breastfeeding, sore breasts, and weight loss due to breastfeeding.

Topic	Count	Name
15	1,141	15_formula_breastfeeding_breastfeed_fed
56	461	56_sore_boobs_breasts_symptoms
196	176	187_weight_breastfeeding_lose_lost
349	94	349_classes_class_breastfeeding_birthing
851	35	851_lamictal_breastfeeding_breastfeed_lam
1004	29	1004_pump_breast_champva_willow
1052	27	1052_aeroflow_contacted_insurance...
1151	26	1115_lump_cancer_breast_lumps

TABLE A2. **BERTopic** identified “breastfeeding” topics in April 2019 birth club.

Initial topic analysis suggested that breastfeeding discussion encompasses more than just feeding concerns and relates to the experience of the breastfeeding person, the impact of breastfeeding on medication usage, and insurance and reimbursement concerns. We then explored the top breastfeeding topic more closely; first, we collected all posts of users who reported struggling breastfeeding in the major breastfeeding topic. Then, we conducted two separate topic modelings; one topic model on all such user posts, and one topic model on all such user posts excluding the specifically identified posts from the breastfeeding topic of the full data set.

Our further exploration of the breastfeeding topic including all posts yielded noticeable differences in top 10 identified topics. Breastfeeding, of course, ranked higher than in the full dataset. Terms for induction, sleep, and percentile

Topic	Count	Name
1	2,867	(omitted and/or missing post data)
2	1,287	2_seat_stroller_car_infant
3	850	3_percentile_lbs_measuring_oz
4	824	4_leave_fmlla_job_disability
5	769	5_induction_induced_inductions_induce
6	717	6_bump_bloat_belly_look
7	712	7_movement_kicks_movements_flutter
8	652	8_sleep_exhausted_tired_nap
9	514	9_formula_breastfeeding_breastfeed_fed
10	637	10_chicken_potatoes_sauce_meals

TABLE A3. BERTopic identified topics in subset of posts of users who self-reported breastfeeding struggles in April 2019 birth club main “Breastfeeding” topic.

Topic	Count	Name
1	2,866	NA (omitted and/or missing post data)
2	843	2_percentile_measuring_lbs_oz
3	850	3_induction_induced_inductions_39
4	824	4_leave_fmlla_job_disability
5	769	5_vaccine_vaccines_flu_shot
6	717	6_bump_bloat_belly_look
7	712	7_sleep_exhausted_tired_nap
8	652	8_boy_girl_gender_boys
9	514	9_stroller_seat_infant_car
10	637	10_throat_cough_cold_fever

TABLE A4. BERTopic identified topics in subset of posts of users who self-reported breastfeeding struggles in April 2019 birth club main “Breastfeeding” topic, excluding the posts in which they self-reported struggles.

weights rated about the same as in the full dataset, which may be related to these items effecting many aspects of pregnancy and post-pregnancy experiences. Interestingly, leave and FMLA rated much more highly in the breastfeeding struggling subset of data, which suggests that this may be a correlated risk factor in early cessation of breastfeeding that should be explored further.

In the topic modeling that excluded the specifics posts that indicated self-report of breastfeeding struggle, we notice many of the same trends as the previous topic modeling. Interestingly, a cold/sore throat/fever topic emerges, which may suggest that being ill (whether it be child or parent) could be another risk factor that should be studied further in relation to breastfeeding.

Additionally, we assessed the relative importance of anxiety and depression labeled topics in the various subsets of data, and noticed that rank and importance of the top anxiety-labeled and depression-labeled topics (that is, in topics that contain the terms anxiety or depression) increased in both subsets of breastfeeding struggles (with and without the coded posts).

Topic tag	Subset	Full Data	Subset without Posts
Anxiety	122, 0.0009	221, 0.0005	117, 0.001
Depression	266, 0.0005	297, 0.0003	198, 0.006

TABLE A5. Comparison of rank and percentage of topic in respective corpus for highest ranked anxiety- and depression-labeled topics in subset of self-reported struggling breastfeeding users topics (topic count: 842), full data set (topic count: 2,120), and dropping all identified self-reported struggling breastfeeding posts (topic count: 854).

Topic	Count	Name
1	11,039	0_name_names_middle_named
2	10,564	1_na_af_been_
3	4,224	2_kicks_movement_movements_flutter
4	4,112	3_measuring_percentile_oz_lbs
5	2,866	4_bump_bloat_showing_bloated
6	2,673	5_spotting_bleeding_brown_clots
7	2,495	6_induction_induced_induce_39
8	1,931	7_sleep_exhausted_nap_tired
9	1,854	8_ultrasound_ultrasounds_3d_elective
10	1,854	9_nausea_sickness_nauseous_sick

TABLE A6. BERTopic identified topics in April 2019 birth club, alternate run.

Topic	Count	Name
153	197	153_breastfeeding_breastfeed_breastfed_wean
187	171	187_formula_fed_breastfeeding_feeding
224	143	224_breastfeeding_weight_lose_lost
232	136	232_poop_pooping_poops_pooped
234	134	234_class_classes_breastfeeding_birthing
331	94	331_awesome_cool_wow_thats
725	41	725_pills_diet_breastfeeding_colace
800	38	800_lamictal_breastfeeding_breastfeed_lam
1256	21	1256_latch_latching_formula_nipple
1733	14	1733_wake_wakes_feed_eats
2027	10	2027_nursing_patpat_tops_blousesshirts
2028	10	2028_nipples_nipple_breastfeeding_painlike

TABLE A7. BERTopic identified “breastfeeding” topics in April 2019 birth club, alternate run.

Classifier	Accuracy
Logistic Regression	0.92
K-Nearest Neighbors	0.92
Support Vector Machine	0.92
MultinomialNB	0.92
Decision Tree	0.89
Random Forest	0.92
Gradient Boosting	0.88
AdaBoost	0.91
Perceptron	0.90
Ridge Classifier	0.90
Nearest Centroid	0.70

TABLE A8. Classifiers’ accuracy on human labeled data ($n = 1,141$ documents, 3 human coders) (80% train/20% test). Using three coders (CB, IS, and PS), we independently annotated the predominant breastfeeding topic in the April 2019 birth club to identify self-report of struggles associated with breastfeeding. Our inclusion criteria centered on explicit mention of physical or mental struggles during breastfeeding (e.g. pain, soreness, frustration, worry) as well as self-report of cessation of breastfeeding before 1 year (based on the 2019 American Academy of Pediatricians recommendation of 1 year duration of exclusive breastfeeding, which changed to 6 months as of 2023 [45]). We excluded anecdotes from users about other people who struggled with breastfeeding (e.g., relating a story about a friend who struggled) and attempt to identify current pregnancy/birth struggles (i.e., non-retrospective about previous births before April 2019). We required all coders to rate a post as “negative” to include the breastfeeding struggle in the subset topic modeling; if coders disagreed, they discussed its inclusion until we reached consensus. Upon initial coding, prior to discussion, coders achieved 83.3% agreement (590/708 posts).

doc_topic_prior (α)	topic_word_prior (η)	number of documents (k)	perplexity score
0.25	0.25	4	2261.9289
0.1	0.1	10	2189.5068
0.05	0.05	20	2135.6854
0.02	0.02	50	2276.6576
0.01	0.01	100	2588.2222
0.005	0.005	200	3192.3299
0.25	0.1	4	2240.2458
0.1	0.1	10	2189.5068
0.05	0.1	20	2179.0988
0.02	0.1	50	2481.3575
0.01	0.1	100	3155.0606
0.005	0.1	200	4751.3159
0.01	0.25	4	2279.6107
0.01	0.1	10	2235.3667
0.01	0.05	20	2208.6584
0.01	0.02	50	2366.8750
0.01	0.01	100	2588.2222
0.01	0.005	200	2934.7749

TABLE A9. LDA hyperparameter combinations and resulting perplexity using `sk-learn` Latent Dirichlet Allocation (LDA) function. The parameter settings that yielded the lowest perplexity score on the full dataset was $\alpha = 0.5, \eta = 0.5, k = 20$.