Facilitating Online Healthcare Support Group Formation using Topic Modeling

Pronob Kumar BARMAN [a,1], Tera L REYNOLDS [a], James FOULDS [a]
[a] *Department of Information Systems, University of Maryland, Baltimore County*
ORCID IDs: Pronob Kumar Barman: 0009-0009-0037-3302, Tera L Reynolds: 0000-0001-7536-5504, James Foulds: 0000-0003-0935-4182

Abstract. Patients increasingly seek peer support in online health forums; however, the large-scale and inconsistent engagement patterns of these forums often fail to meet patients' support needs effectively. Smaller, personalized support groups could address these challenges by tailoring interactions to users' shared experiences and demographics. This study introduces the Group-specific Dirichlet Multinomial Regression (gDMR) model, a structured framework for automating and personalizing support group formation using user-generated content, demographic, and interaction data. By incorporating group-specific parameters and node embeddings, gDMR captures nuanced demographic and behavioral patterns, extending traditional Dirichlet Multinomial Regression (DMR). Experiments demonstrate gDMR's ability to form groups that are more semantically coherent and contextually relevant than those produced by baseline models. This scalable model reduces manual effort in group personalization, fostering inclusive engagement and enhancing patient-centered care. Findings highlight gDMR's potential as a framework for digital healthcare platforms, from online health communities to healthcare systems utilizing electronic health records (EHRs), advancing health informatics through support group formation.

Keywords. Healthcare support groups, Topic modeling, Online health forums, Personalized support groups

1. Introduction

Patients increasingly rely on online health forums to share experiences, seek advice, and find emotional support throughout their healthcare journeys. Research has shown that peer support within these communities can significantly improve patient outcomes [1,2,5,11]. However, forming cohesive support groups presents both technical and social challenges, particularly in balancing diverse user demographics, interests, and interaction patterns. Manual group formation is labor-intensive, lacks personalization, and can result in inconsistent engagement [8,7].

To address these challenges, we introduce the *Group-specific Dirichlet Multinomial Regression* (gDMR) model, an automated framework for personalized support group formation. This model synthesizes user-generated content, demographics, and interaction data to form groups that reflect shared themes and experiences. By incorporating

[1]Corresponding Author: Pronob Kumar Barman, Department of Information Systems, University of Maryland, Baltimore County; E-mail: pbarman1@umbc.edu.

group-specific parameters and node embeddings, gDMR models the relationships among users, creating smaller, more personalized groups that are responsive to user interactions. Personalized support groups have the potential to make online health communities more inclusive and effective by addressing patients' unique social and psychological needs.

Through experiments, we demonstrate that gDMR with node embeddings significantly outperforms baseline models, such as LDA [4] and DMR [3], in perplexity, held-out log-likelihood, and topic coherence. Node embeddings improve topic coherence, enabling the formation of semantically cohesive groups that promote meaningful interactions. This scalable framework reduces manual effort, offering an effective solution for enhancing engagement and personalization in online health communities.

2. Methods

The dataset was collected from **MedHelp**, an online health community that closed on May 31, 2024. It includes over 2 million questions and 8 million answers from users discussing topics such as chronic diseases (e.g., diabetes, asthma) and general wellness (e.g., fitness, diet) [12]. Despite the platform's closure, the dataset offers useful examples for studying support group formation and may inform applications in other online health community or even electronic health record (EHR) systems to support local support group formation (in-peron, online, or hybrid).

We performed the following steps to address missing data values: 1. *Country* was encoded as "Unknown." 2. *Age* was replaced with the mean. 3. *Gender* was left unspecified if undetermined. 4. *Membership year* was imputed using the mode.

Model Overview Our proposed gDMR model leverages and extends the DMR model [3], tailoring it to the specifics of initial group formations based on user-generated content and demographic features. The model encodes the intuition that users have mixed membership over idealized groups, and that both their text and their demographics contribute to these memberships. It takes as input a predefined number of groups G, a Dirichlet hyperparameter β (represented as a vector of length V, where V is the vocabulary size), and for each user u, text sequences w_u (of length W_u) and user-specific demographic feature vectors $x_u \in \mathbb{R}^D$ (where D is the number of demographic variables). The learning algorithm employs Gibbs sampling to estimate posterior distributions, producing idealized group membership distributions θ_u, and for each group, word distributions ϕ_g and regression coefficients $\lambda_g \in \mathbb{R}^D$. The pseudocode for the model's assumed generative process is:

1. For each group g from 1 to G:

$\phi_g \sim \text{Dirichlet}(\beta)$ \\Word distributions per group

$\gamma_g \sim \text{Gamma}(1,1)$ \\Prior pseudocounts per group

$\lambda_g \sim \text{N}(\mu, \sigma^2 I)$ \\Regression coefficients

2. For each user u from 1 to U:

$$\alpha_{ug} = \exp((x_u)^T \lambda_g) + \gamma_g$$
$$\forall g \in \{1, \ldots, G\}$$
$$\theta_u \sim \text{Dirichlet}(\alpha_u)$$

\\Group distributions per user

3. For each word w from 1 to W_u:

$z_{uw} \sim \text{Categorical}(\theta_u)$ \\Select group

$w \sim \text{Categorical}(\phi_{z_{uw}})$ \\Select word

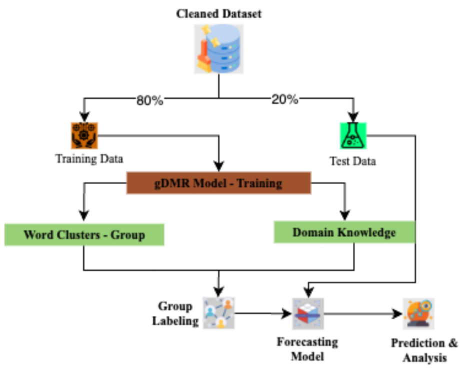

Figure 1. Model Methodology Flowchart.

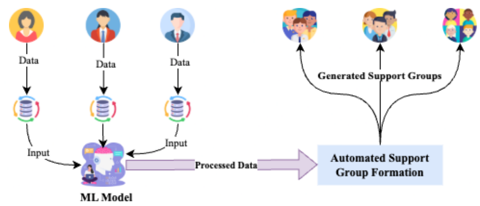

Figure 2. Support Group Formation Flowchart.

We set hyperparameters to $G = 20$, $\beta = 0.01$, $\sigma = 1.0$, with 1000 Gibbs iterations, 350 burn-in, and regression updates after 700 iterations (leveraging first 700 LDA iterations to ensure the establishment of initial coherent groups). The final 300 iterations combined collapsed Gibbs sampling and BFGS optimization every 10 Gibbs iterations. Figures 1 and 2 depict our training procedure steps and model overview of support group formation, respectively.

Our gDMR model incorporates node embeddings derived from user interaction data to capture latent relationships through graph-based data, enhancing group formation by adding structural interaction features alongside demographics and text. We construct a directed graph where the nodes represent users and the edges indicate interactions (e.g., replies or comments), weighted by frequency and tagged with `post_id` for context. Using Node2Vec [9], we generate 64-dimensional embeddings via 200 random walks per node, capturing both local and global network structures. In practice, this means that users who interact more frequently within the community are likely to be grouped together, reflecting shared interests or themes. This approach leverages interaction patterns to enrich gDMR's feature set, improving the contextual relevance of the generated support groups.

3. Results

Our experimental results demonstrate that the gDMR model with node embeddings outperforms baseline models, such as LDA and DMR, across held-out log-likelihood, topic coherence, and perplexity, underscoring its superior predictive accuracy, interpretability, and group cohesion for support group formation. As shown in Table 1, the gDMR model with embeddings achieved the highest held-out log-likelihood score (-403.350), surpassing gDMR without embeddings (-463.778), LDA (-606.874), and DMR (-465.230), emphasizing the value of interaction data for generalization. Topic coherence scores, also presented in Table 1, highlight that gDMR with embeddings (-2.401) outperformed other models, such as LDA (-4.488), DMR (-4.789), and gDMR without embeddings (-5.048), which indicates that embeddings contribute to more interpretable and thematically consistent topics. Figure 4 illustrates that within-group similarities showed a higher median and broader spread than the random baseline, confirming the model's effectiveness in creating semantically cohesive groups essential for meaningful user engagement. Finally,

Figure 3. Perplexity comparison between gDMR (with and without node embeddings) and DMR models across iterations (truncated at the point of minimum perplexity).

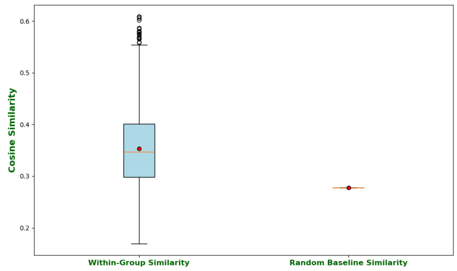

Figure 4. Comparison of Within-Group and Random Baseline Similarities, highlighting the enhanced semantic coherence within groups.

Model	Log-Likelihood	Coherence Score
LDA Model	-606.874	-4.488
DMR Model	-465.230	-4.789
gDMR Model (without node embeddings)	-463.778	-5.048
gDMR Model (with node embeddings)	-403.350	-2.401

Table 1. Comparison of Log-Likelihood and Coherence Scores across Models

as shown in Figure 3, perplexity, which measures how well a model predicts a sample of data, decreased sharply in the gDMR model with embeddings. Lower perplexity values indicate better generalization performance, as they suggest the model is assigning higher probabilities to the observed data. The gDMR model with embeddings achieved the lowest perplexity value (1,600), with a 52.2% improvement over DMR (3,350) and a 38.5% improvement over gDMR without embeddings (2,600), further underscoring the model's superior generalization capability and ability to form coherent support groups.

Support Group Formation Our approach used text, user features, and gDMR-generated group membership distributions (Figure 1 and 2) based on demographics, node embeddings, and content similarity, further refined into cohesive support groups using thematic and feature similarity measures. Thematic similarity was calculated via TF-IDF vectorization between user posts and group profiles, weighted at 70%, while feature similarity, weighted at 30%, incorporated user embeddings and demographics. Constrained K-Means clustering further refined these groups, balancing sizes by merging clusters with fewer than 10 members and splitting those with more than 22.

To improve interpretability, groups were made country-specific. Example groups include: 1. **Russia**: **Digestive health**, keywords: "stomach," "diet" (18 members). 2. **Turkey**: **Cardiovascular health**, keywords: "blood," "pressure," "cholesterol" (22 members). 3. **Southeast Asia**: **Liver health**, keywords: "liver," "enzyme levels" (15 members).

As shown in the above examples and by our experimental results, combining gDMR with constrained K-Means clustering allowed us to create demographically and thematically relevant groups. Node embeddings captured interaction patterns, enhancing the model's ability to support meaningful and contextually appropriate group assignments.

4. Discussion

Our gDMR model facilitates personalized support group formation by leveraging user-generated content, demographics, and interaction patterns. It can potentially integrate with clinical systems, such as Electronic Health Records (EHRs), enabling healthcare providers to automate and personalize patient support, though real-world validation remains an essential next step [10]. Evaluating gDMR's long-term effects on patient engagement, adherence, and overall health outcomes will be necessary to fully understand its clinical applicability and sustainability. Algorithmic bias remains a critical ethical concern; demographic and interaction-based group assignments risk marginalizing underrepresented users or reinforcing existing disparities. To mitigate these risks, incorporating fairness-aware learning techniques and continuously monitoring outcomes across demographic groups will be imperative. Comprehensive discussions of these ethical considerations, as well as detailed evaluations of real-world applicability and long-term impacts, will be explored extensively in our future research.

5. Conclusions

Our gDMR model provides a scalable solution for forming personalized support groups, addressing limitations of traditional methods in online health communities. Experimental results show superior predictive accuracy and topic coherence, while effectively leveraging user content and interaction data. Future work will evaluate real-world impact, explore adaptive group assignments, and implement bias safeguards, ensuring gDMR promotes equitable, culturally sensitive, and ethical practices in patient-centered care.

References

[1] Worrall H et al. Effectiveness of support groups: literature review. Ment Health Soc Incl. 2018;22(2):85–93. doi:10.1108/MHSI-12-2017-0055.

[2] Mayo Clinic. Support groups: make connections, get help [Internet]. 2023 [cited 2025 Mar 14]. Available: https://www.mayoclinic.org/healthy-lifestyle/support-groups/in-depth/support-groups/art-20044655

[3] Mimno D, McCallum A. Topic models conditioned on arbitrary features. arXiv:1206.3278; 2012.

[4] Blei DM et al. Latent Dirichlet Allocation. J Mach Learn Res. 2003;3:993–1022.

[5] Naslund JA et al. Future of mental healthcare: peer-to-peer support and social media. Epidemiol Psychiatr Sci. 2016;25(2):113–22. doi:10.1017/S2045796015001067.

[6] Wallach HM et al. Evaluation methods for topic models. Proc ICML. ACM; 2009. p.1105–12.

[7] Müller AM et al. Group formation: distribution of extraversion in educational settings. Acta Psychol (Amst). 2024;242:104111. doi:10.1016/j.actpsy.2023.104111.

[8] Bergtold JS, Shanoyan A. Group formation methods and group-based learning performance. Front Educ. 2024;9:1362211. doi:10.3389/feduc.2024.1362211.

[9] Grover A, Leskovec J. Node2Vec: scalable feature learning for networks. Proc ACM SIGKDD. ACM; 2016. p.855–64. doi:10.1145/2939672.2939754.

[10] Tshimula JM et al. Redesigning EHR systems for developing countries. Proc Int Conf Med Health Inform. ACM; 2023. p.216–21. doi:10.1145/3609791.3609825.

[11] Cha BS et al. Integration of peers in multi-site digital mental health innovation. Adm Policy Ment Health. 2024;51(2):226–39. doi:10.1007/s10488-023-01331-5.

[12] Vydiswaran VGV, Reddy M. Identifying peer experts in online health forums. BMC Med Inform Decis Mak. 2019;19(Suppl 3):68. doi:10.1186/s12911-019-0782-3.