Identifying Loneliness Themes in University Reddit Communities via Topic Modeling

1st Jenny Ye
School of Information
University of Michigan
Ann Arbor MI, United States
jennyye@umich.edu

***Abstract*—We mined 2024–25 posts from *r/college* and *r/GradSchool* to map how students discuss loneliness. Three models—LDA, Sentence-BERT + K-Means, and BERTopic—progressively exposed broad, overlapping, and fine-grained themes. LDA outlined five macro topics; embedding clusters showed their semantic overlap; BERTopic split them into 24 niches, including orientation-week isolation, dorm-room conflicts, and PhD advisor stress. These insights can guide targeted campus mental-health outreach.**

***Index Terms*—Sentiment Analysis, LDA, Sentence-BERT, BERTopic, Loneliness**

I. Introduction

College-student loneliness surged after COVID-19, with surveys showing over half now feel isolated. Reddit mirrors this trend: thousands of posts in *r/college* and *r/GradSchool* chronicle social anxiety, friend-making hurdles, dorm solitude, and lab isolation. To distill those narratives we analyzed all 2024–25 posts containing "lonely," "alone," "isolated," or "social anxiety" through a three-stage pipeline: (1) LDA for a five-topic macro map, (2) Sentence-BERT + K-Means to expose semantic overlaps, and (3) BERTopic to surface 24 fine-grained subthemes—from orientation-week loneliness to PhD-advisor stress. Recent work has moved beyond bag-of-words models by pairing transformer embeddings with density-based clustering, as in BERTopic [4], and by fusing emotion-labeled corpora such as GoEmotions [5] with mental-health classifiers. Large-scale Reddit studies now track temporal shifts in anxiety and depression using these neural topic models, establishing a state-of-the-art toolkit that we extend to the specific lens of student loneliness.

II. Method

A. Data Collection and Preprocessing

Using the PullPush.io mirror of Pushshift, we scraped several thousand posts and top-level comments (Jan 2024–Mar 2025) from *r/college* and *r/GradSchool* that contained "lonely," "alone," "isolated," or "social anxiety." After lower-casing text, stripping URLs, punctuation, digits, and a campus-specific stop-word list, we kept tokens intact (no stemming) to preserve nuances such as "lonely" versus "alone." The cleaned text fed transformer models directly, while its stop-word-filtered version served the bag-of-words LDA.

B. Topic Modeling Approaches

We applied three complementary topic-modeling pipelines.

1) *LDA (Gensim):* First, a 5-topic Latent Dirichlet Allocation (LDA, Gensim) on the bag-of-words corpus provided a macro-level map of recurring themes, with the number of topics chosen by coherence and readability.
2) *Sentence-BERT + KMeans:* we generated 384-D sentence embeddings with the pre-trained all-MiniLM-L6-v2 Sentence-BERT model using Hugging Face and clustered them using K-Means ($k = 5$ gave the clearest separation). A t-SNE projection visualised these semantic clusters. [3]. These 384-dimensional embeddings capture semantic similarity between posts. We then performed KMeans clustering on the embeddings to group posts into k clusters (we experimented with $k = 5$ to $k = 7$ and found $k = 5$ provided distinct, meaningful clusters).
3) *BERTopic:* We applied BERTopic [4] to the same set of Sentence-BERT embeddings. BERTopic uses a class-based TF-IDF (c-TF-IDF) to derive topic words for clusters found in the embedding space (using HDBSCAN for density-based clustering). the same embeddings were fed to BERTopic, which pairs density-based HDBSCAN with class-TF-IDF; a custom stop-word list removed generic Reddit terms and the query keywords themselves. BERTopic automatically produced two-dozen fine-grained topics, merging near-duplicates to maximise coherence.

III. Results and Analysis

A. LDA Topics

A 5-topic LDA model gives a high-level map of student loneliness. Topic 1 groups "friends, meet, party," reflecting difficulty making friends; Topic 2 ("anxiety, shy, awkward") captures social-anxiety posts; Topic 3 ("dorm, roommate, weekends") points to isolation in campus housing; Topic 4 ("grad, research, advisor") isolates graduate-research solitude; Topic 5 ("home, family, remote") signals homesickness after

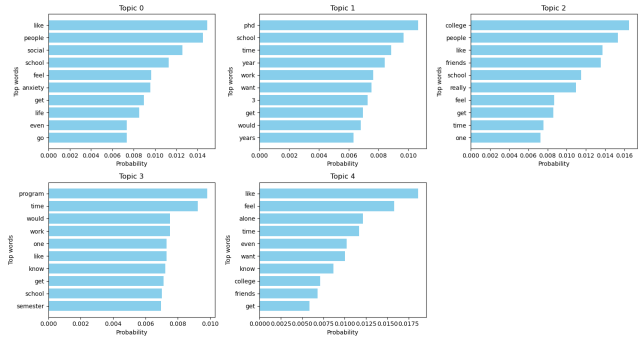

Fig. 1. LDA results: Top words for each of the 5 discovered topics on the Reddit corpus. The topics broadly correspond to (1) friendship-making struggles, (2) social anxiety, (3) dorm loneliness, (4) grad school isolation, and (5) homesickness/remote life.

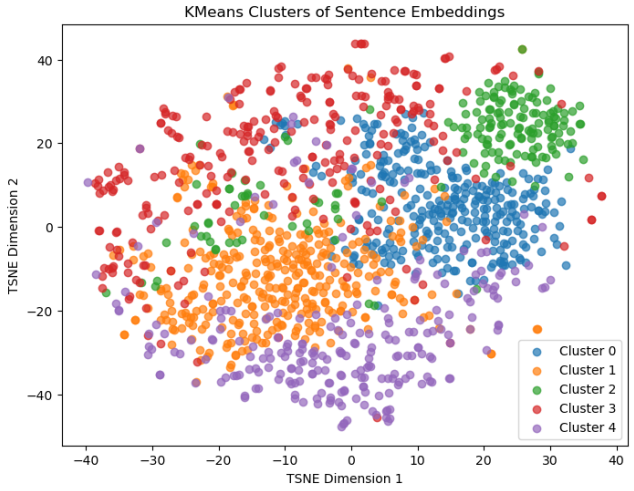

Fig. 2. t-SNE visualization of Sentence-BERT embeddings of Reddit posts, with points colored by KMeans cluster (5 clusters). Each cluster represents a different loneliness theme (e.g., friendship difficulties, social anxiety, grad student isolation, roommate issues, homesickness).

remote learning. Together these five buckets cover the dominant ways students describe feeling alone (Fig.1).

B. Sentence-BERT+K-Means—Cluster View

Clustering 384-D Sentence-BERT embeddings (k! =!5) mirrors LDA themes while exposing their overlaps (Fig.2). Clusters align with (1) freshman friend-making woes, (2) social-anxiety and fear of rejection, (3) PhD/grad-lab isolation, (4) roommate conflicts in dorms, and (5) homesickness—yet the t-SNE plot shows partial blending between the friend-making and social-anxiety clusters, illustrating how those experiences often co-occur.

C. BERTopic—24 Fine-Grained Topics

BERTopic's HDBSCAN clustering yielded 24 fine-grained topics (Fig.3), subdividing broad LDA themes into concrete scenarios—e.g., orientation-week loneliness, class-presentation anxiety, PhD-advisor stress, financial-aid worries,

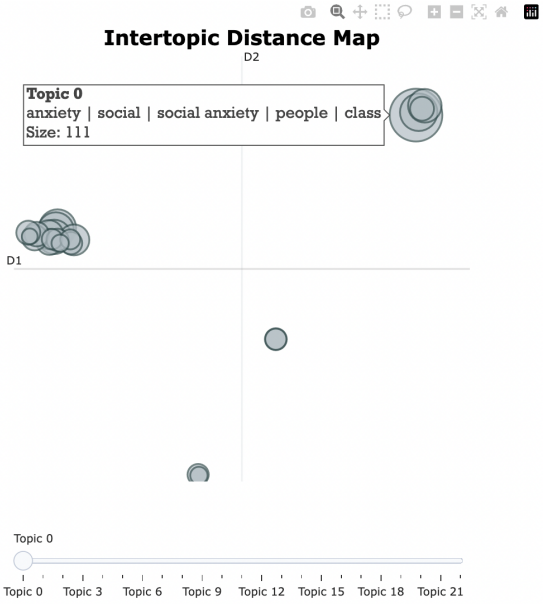

Fig. 3. BERTopic results: Distribution of 24 fine-grained topics related to loneliness. Each bar corresponds to a topic, labeled by its key theme (e.g., roommate issues, orientation loneliness, social anxiety in class, academic burnout, etc.). Topics are ordered by frequency of posts.

and ADHD-related burnout. These nuanced clusters underscore the many faces of campus loneliness and point to targeted remedies such as orientation social events or advisor-mentor programs for grad students.

IV. Conclusion

Our three-model pipeline paints a layered picture of student loneliness on Reddit. LDA supplies the macro view—five recurring themes such as social anxiety, friend-making struggles, dorm isolation, and grad-school solitude. Sentence-BERT + K-Means confirms these themes and visualizes how some overlap (e.g., social anxiety and friendship worries), while BERTopic drills down to 24 niche scenarios—from orientation-week loneliness and dining-hall panic to PhD-advisor stress—revealing the nuanced situations that feed isolation.

These insights can guide campus support. Orientation programs could add peer-bonding events, graduate departments might expand mentoring, and housing offices could mediate roommate conflicts. Monitoring such online discourse—and enriching it with future sentiment analysis—offers universities a real-time barometer of student well-being and a data-driven basis for targeted interventions.

References

[1] Sodexo, "More Than 50% of Gen Z College Students Report Feeling Lonely," Newsroom Press Release, Aug. 2022.
[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. of EMNLP*, 2019.
[4] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv:2203.05794*, 2022.

[5] D. Demszky *et al.*, "GoEmotions: A Dataset of Fine-Grained Emotions," *arXiv:2005.00547*, 2020.
[6] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The Pushshift Reddit Dataset," in *Proc. of ICWSM*, vol. 14, pp. 830–839, 2020.