





Analyzing Collective Intelligence Through Sentiment Networks in Self-organized Douban Communities

Tiantian Xie¹  and Xiaokun Wu^{1,2} 

¹ School of Journalism and Communication, South China University of Technology, Guangzhou 510006, Guangdong, China

wuxiaokun@scut.edu.cn

² Center for Data Analysis and Information Visualization, South China University of Technology, Guangzhou 510006, Guangdong, China

Abstract. Understanding the communication behaviors of online community users, particularly their crowd intelligence dynamics, has long been a focal point within network communication research. In this study, we present an approach integrating the BERTopic topic model, advanced Natural Language Processing (NLP) techniques, and Social Network Analysis, to meticulously dissect the intricate dynamics of emotion propagation and evolution within collective behaviors that unfold on social networks. Our investigation delves deeply into this complex landscape, exploring the relationships between the sentiment of the initial post and subsequent responses, the interplay between sentiment strength and activity levels, and the correlation between sentiment polarity and the intensity of activity. This study highlights the significance of harnessing the combined power of BERTopic, NLP, and social network methodologies to decode the subtleties of emotional propagation and transformation.

Keywords: Collective intelligence · BERTopic · Social Network Analysis · Natural Language Processing · Sentiment distribution

1 Introduction

Social anxiety disorder, often called “social fear,” and its counterpart, the “social master” or “social cow,” represent contrasting social traits in society. Social fear involves nervousness, discomfort, and social avoidance, while social masters excel in and enjoy socializing. In the Douban virtual community, these concepts offer an excellent opportunity for comparison. Both social fear and social cow groups in virtual communities share similarities. The former consists of individuals seeking support for social unease, while the latter includes those who thrive in social activities. This contrast forms a valuable framework for research, shedding light on community interaction behaviors. Through a careful comparative analysis, we gain insights into the emotional expressions and evolution of these traits in a networked environment. This research focuses on Douban, China’s largest online

interest-based community, with approximately 120,000 users, mainly aged 19 to 30. The Douban mobile app has around 12 million monthly and 3 million daily active users. We employ Bert topic modeling and semantic network analysis to construct comment interaction networks, uncovering virtual community dynamics. We analyze member contributions using a sentiment dictionary and conduct regression analysis to explore emotional changes during community engagement. The key contributions are as follows:

- By harnessing social network data from online community users, an opportunity arises to gain profound insights into the diffusion patterns and driving factors that underpin information flow within social networks. This fills a crucial void left by traditional social survey research, which often struggles to establish contact with specific demographic groups and comprehend the intricacies of their collective intelligence behaviors.
- Scrutinizing the intricacies of information dissemination within a particular community not only facilitates a grasp of the interconnections, patterns of association, and routes taken by information among group members, but also provides invaluable insights applicable to broader fields of social network research. These analytical outcomes serve as a point of reference for diverse areas of study.
- Social terror groups often coalesce into networks of mutual aid and solidarity within virtual communities. Delving into the dynamics of interactions and supportive bonds among community members unravels the intricacies of the social support network, affording comprehension of its evolution, composition, and influential factors. This analytical lens offers a foundation for investigating topics encompassing community dynamics, advocacy, mental well-being, and beyond.

2 Related Works

Previous research on social anxiety has largely focused on medical aspects, including cultural influences [9], educational backgrounds [6], social comparisons [11, 13], and the psychological consequences [5]. However, these studies often overlook the role of information interactions. To address this gap, our study collects data from two control communities and uses methods like social network analysis and topic modeling to uncover how information exchange and emotional convergence occur in virtual networks. We also examine factors shaping collective mentality in information dissemination. Our main goal is to reveal the mechanisms of swarm intelligence behavior in specific social contexts, an area less explored in existing studies.

To investigate collective intelligence in virtual communities, we utilize diverse tools and advanced technologies, including BERTopic, social network analysis, and sentiment analysis. Our objective is to comprehensively understand the emotions and behaviors within these virtual communities.

2.1 BERTopic Topic Modeling Method

BERTopic serves as a sophisticated topic modeling technique that extracts coherent topic representations by introducing a category-based variant of TF-IDF. It initiates the process by generating embeddings through a pre-trained model. Subsequently, the Unified Manifold Approximation and Projection (UMAP) methodology is applied to reduce the dimensionality of these embeddings. This reduction enhances the efficiency of document clustering. To finalize this clustering, the HDBSCAN algorithm, rooted in hierarchy and density-based principles [7], is employed.

UMAP is a nonlinear dimensionality reduction method that preserves local and global data relationships by creating a data domain, topology and refining low-dimensional representations. HDBSCAN is a density clustering algorithm for grouping data points into interconnected clusters while identifying varying densities and noise points. UMAP and HDBSCAN are prominent algorithms for data analysis and clustering. UMAP maps high-dimensional data to lower dimensions using topological insights and distance relationships. HDBSCAN organizes data points into closely related clusters, excelling with diverse datasets due to its adaptability to varying densities and noise.

BERTopic offers significant advantages over other topic modeling approaches by bridging the gap between density-based and center-based sampling methods. It excels in versatile data scenarios due to its unique ability to combine these strategies, enhancing its utility [2]. Furthermore, BERTopic's adaptability extends to various language models, enhancing its performance through co-optimization. This optimization improves accuracy and efficiency in topic modeling, as confirmed by studies [2]. Studies substantiate BERTopic's exceptional performance in language processing, increasing the General Language Understanding Evaluation (GLUE) score to an impressive 80.5%, reflecting a substantial improvement of 7.7% [2]. It outperforms alternatives by using pre-trained models, eliminating the need for data labeling and yielding context-infused embeddings for words and sentences, particularly beneficial for short text data [2].

2.2 User Sentiment Analysis Within Community Texts

Currently, sentiment analysis techniques within online communities can be broadly categorized into three main approaches: keyword-based sentiment analysis, machine learning-based sentiment analysis, and hybrid sentiment analysis techniques [15].

Keyword-based methods use dictionaries, requiring text preprocessing for accuracy. Sentiment words are extracted manually or from predefined dictionaries. Construction of these dictionaries is challenging and can incur costs [17]. Moreover, different types of texts demand distinct dictionaries, as employing the same dictionary for diverse text analyses may lead to various errors. Nonetheless, as artificially constructed dictionaries continue to improve, the accuracy of keyword-based sentiment analysis technology is expected to rise gradually.

Machine learning methods use labeled/unlabeled data and algorithms like SVM, decision trees, KNN, and Bayesian networks. They offer flexibility in choosing text features and classifiers for emotional text analysis. For instance, Md. Rafiqul Islam et al. used decision trees, SVM, KNN, and ensemble methods to analyze depression-related Facebook data, covering emotional characteristics, temporal attributes, and language style [10]. MICHAEL M. TADESSE et al. combined natural language processing and machine learning to uncover depression-related features in Reddit posts. SVM achieved 80% accuracy in extracting individual features, while the Multi-Layer Perceptron (MLP) excelled, reaching 91% accuracy with LIWC+LDA+bigram [20]. IGonzalo A. Ruz et al. used Bayesian networks to explore connections among sentiment words in Spanish datasets, enhancing sentiment analysis understanding [18].

In text sentiment analysis, neural network methods include single networks, hybrid models, attention-based approaches, and pre-trained models. They learn complex text features and context, improving understanding. For example, Zhou Yu et al. used a convolutional neural network with a Boltzmann machine for latent topic modeling and sentiment analysis [22]. However, deep learning methods require substantial data and time, limiting their effectiveness with smaller datasets and large-scale analysis.

Given the unique nature of our research sample, this study primarily employs sentiment analysis techniques based on sentiment dictionaries. Utilizing these artificially crafted lexicons, we compute the frequency of sentiment-infused words within the text and weigh them against other words to derive a sentiment index. It's worth noting that distinct texts necessitate the use of specific dictionaries, as applying a uniform dictionary to disparate texts may introduce varying errors. Nonetheless, as artificially constructed dictionaries continue to evolve and refine over time, the accuracy of keyword-based sentiment analysis technology has shown a progressive improvement.

3 Data and Methods

3.1 Data Collection

This study collects data from two virtual community groups on the Douban platform: the “Social Ability Rehabilitation Group” (Group 1) and the “Social OCD Group” (Group 2). Group 1 data was collected from January 17, 2021, to November 22, 2021, including 4,875 valid posts and 54,398 comments. Group 2 data covers September 7, 2021, to June 12, 2022, with 1,252 posts and 13,546 comments. Both groups were observed for approximately 10 months from their inception. This approach provides a comprehensive view of group dynamics and ample data for sentiment and social network analysis.

The dataset includes poster ID, post content, timestamps, responder IDs, reply content, and reply timestamps, forming a robust foundation for in-depth analysis of social interactions and emotional expressions.

3.2 Data Preprocessing

In this paper, we ensure data integrity, accuracy, and consistency. We first inspected crawled data for attributes and temporal gaps. We then conducted preprocessing, fixing missing, duplicate, and aberrant posts. Finally, we normalized the data for standardized format. These efforts yielded: Group 1 with 4,868 validated content posts and 54,120 comments, and Group 2 with 1,243 content posts and 15,450 comments.

To assess sentiment analysis precision, we sampled and manually annotated 200 entries from both groups. Two trained postgraduate students marked sentiment polarity as 1 for positive, -1 for negative, and 0 for no discernible sentiment. We compared and validated the labels, calculating accuracy (see Table 1).

Table 1. Accuracy of Dictionary-based Sentiment Evaluation

Data set	Posts of G1	Comments of G1	Posts of G2	Comments of G2
Accuracy	0.92	0.94	0.88	0.89

3.3 Data Analysis

Topic Model Analysis. In this study, we employed the BERTopic pre-training model for topic modeling analysis of the textual data. Our methodology involved several key steps:

- Stop Word Compilation: Initially, we merged three separate stop word lists, namely the Harbin Institute of Technology stop word list, Baidu stop word list, and a Chinese stop word list, into a single consolidated stop word list for subsequent processing.
- Text Segmentation and Stop Word Removal: We utilized the jieba tokenizer to segment the text and effectively remove stop words. Additionally, to preserve specific fixed words, we excluded them from the word segmentation process.
- Dimensionality Reduction: Next, we employed the Unified Manifold Approximation and Projection (UMAP) technique to reduce the dimensionality of the word-segmented sentence vectors. This dimension reduction process aids in optimizing the embedding.
- Hierarchical and Density-Based Clustering: Subsequently, we applied the HDBSCAN algorithm, which is rooted in hierarchical and density-based principles, to perform clustering. This step enabled us to identify topics and the key feature words associated with these topics.
- Model Refinement: We made necessary adjustments and optimizations based on the modeling results to enhance the quality of our findings.

- Visualization and Output: Finally, we visualized and output the generated results, presenting the identified topics and their corresponding feature words through visualization techniques.

This comprehensive approach allowed us to effectively analyze and extract meaningful topics and insights from the text data under investigation.

Methods for optimal topic number determination in topic modeling include perplexity and topic similarity assessment. Perplexity, a language model quality measure, is inversely related to test set probability normalized by word count. Lower perplexity signifies better model performance. Typically, perplexity is used to compare topic modeling performance across different topic numbers in the same dataset [4]. Topic similarity evaluation gauges overall topic space distinctiveness and stability by measuring deviations between individual topics and their mean. Greater differences indicate higher topic distinctiveness and overall stability [16].

In our research, we combined perplexity curves with topic distinctiveness analysis. We compared the topic feature words when considering 5, 6, 7, 8, 9, and 10 topics. Eventually, we determined that the optimal number of topics was 8, as it maximized topic distinctiveness and allowed for clear description of the central themes through the topic feature words.

Semantic Network Construction. In text analysis, we employ BERT’s methodology to construct semantic vectors for text content. Using the Transformer encoder unit and multi-head self-attention mechanism, we process and derive vector representations. These vectors undergo further processing through residual connections and normalization layers, followed by a feed-forward and residual network to extract semantic features.

For latent space representation, we use the K-means clustering algorithm to group semantically similar words without supervision. K-means initializes cluster centers based on the desired number of clusters (k) and iteratively minimizes the sum of squared errors, measuring distances between samples and cluster centers, as shown in formula (1).

$$J(c, u) = \sum_{i=1}^M \|x_i - \mu_{c_i}\|^2 \quad (1)$$

where x_i is the i sample, c_i is the cluster of x_i and μ_{c_i} is the center of cluster x_i , M is the total number of samples.

In this study, we use the Python-based TF-IDF algorithm to process main posts and comments in the clustered dataset. TF-IDF helps extract important keywords within each cluster, enhancing our understanding. The extracted data is then visualized and analyzed using Gephi software, creating informative maps that illustrate word connections and associations. These visual representations aid in in-depth analysis of word relationships within the textual data.

Sentiment Communication Analysis. To understand the spread of social anxiety online, we conducted emotion discrimination and contagion analyses on data from Group 1 and Group 2, using an emotion dictionary. Here are the steps:

- Construction of Sentiment Lexicon: We created a sentiment lexicon for supervised learning, including sentiment, negative, and degree words. We expanded an existing Chinese sentiment polarity dictionary with 200 additional emotional words.
- Processing of Main Posts and Texts: We systematically processed primary posts and text data to identify emotional words and compute their scores.
- Degree Word Consideration: We checked if the word before an emotional word was a degree word. If yes, we adjusted the emotional word’s score accordingly.
- Negative Word Assessment: We examined if the word before an emotional word was negative. If an odd number of negative words were found, we multiplied the emotional word’s score by -1. For a double negative, we multiplied by 1.
- Cumulative Score Calculation: We summed emotional word scores across all clauses in the post content to get the post’s overall emotional score.

This comprehensive methodology allowed us to assess the emotional content of posts and gain insights into the propagation of social anxiety in the virtual community.

As a metric for gauging the impact of sentiment contagion within the network, we employ emotional shifts following the reception of information. Our approach involves computing the emotional score for each text by subtracting the negative emotional score from the positive emotional score. The sentiment change is calculated using the following formula:

$$S = S_P - S_N \quad (2)$$

where S_P is the positive emotion score while S_N represents the negative emotion score.

For measuring emotional transmission, we draw inspiration from Michela et al.’s research in Nature [1]. Their study explores how user sentiment and participation behavior transform community structure. We adopt a group dynamics perspective to quantify the interaction between user emotions and engagement. This involves analyzing changes in user emotional values during participation, assessing emotional transformation efficiency, and studying user emotional polarization. In alignment with Michela et al.’s approach, we have redefined three key indicators. These include:

Average Member Sentiment Value ($\sigma(i)$): This metric captures the average emotional disposition of a given community member i , where i represents the i th member, and σ_i denotes the average emotional value derived from all posts and comments authored by member i .

Average Positive/Negative Comment Difference $n\rho(i)$: To compute this indicator, we assess the average disparity between positive and negative comments authored by member i .

Average User Sentiment Transformation($\rho(i)$): This indicator characterizes the emotional changes experienced by virtual community member i during their engagement in the community.

For the calculation of the average positive/negative review difference, we employ the following formula:

$$n\rho(i) = \frac{1}{T_i} \sum_{j=1}^{T_i} (Pos_j(i) - Neg_j(i)) \quad (3)$$

T_i is the number of days the user is active, $Neg_j(i)$ is the number of negative comments and posts issued by the user on day j , $Pos_j(i)$ is the number of positive comments and posts issued by the user on day j .

The formula for calculating the average user emotional polarization is expressed as:

$$\rho\sigma(i) = \frac{((N_i - 2k_i - h_i)(N_i - h_i))}{N_j^2} \quad (4)$$

where N_i , k_i , and h_i respectively refer to the number of total comments and posts, the number of negative comments and posts, and the number of neutral comments and posts.

4 Experimental Results

4.1 Results of BERTopic Topic Modeling Analysis

Using BERT and topic modeling, our aim is to identify topics in text data and group similar texts into categories. This helps us understand and structure extensive text data, uncover hidden topic patterns, and improve text analysis efficiency. We determine the number of topics based on perplexity and topic differentiation assessments, then cluster word vectors from the text accordingly. We reduce the dimensionality of these vectors for two-dimensional visualization. Our observation reveals that the clustering is clearest when using 8 clusters, as shown in Fig. 1. We applied BERTopic topic modeling to the combined text data from Group 1 and Group 2, revealing 8 distinct themes in each group. We thoroughly examined the results and present the topics and explanations for Group 1 and Group 2 in Table 2 below:

Distinctive patterns emerge in both Group 1 and Group 2. Group 1 focuses on interpersonal interaction, emotional support, and self-improvement. They emphasize empathy, problem-solving, and self-discovery. In contrast, Group 2 shares a similar emphasis on interpersonal engagement but leans towards empowerment, social interaction, and relationship dynamics, especially concerning opposite-sex interactions, reflecting their aspirations and relationship expectations.

To assess the comparability and distribution of topics across two virtual communities, we fed the sample data into the BERT model, resulting in the acquisition of topic word embeddings for each topic. To gauge topic similarity,

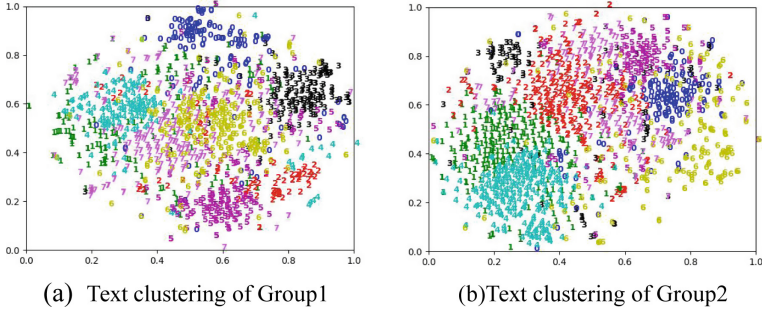


Fig. 1. Text clustering results divided by modules

Table 2. Induction of BERTopic Outputs

Topic model induction for G1	Topic model induction for G2
1) Interpersonal: involves interacting and communicating with other people	1) Interpersonal: involves interacting and socializing with other people
2) Blessing words: including blessings and good wishes	2) Contact with the opposite sex: Communication and interaction with the opposite sex
3) Emotional support: emotional support and comfort	3) Expressing expectations: deals with the topic of expressing expectations and expectations
4) Activities of Daily Living: Topics related to activities of daily living	4) Encouraging Others: Topics related to encouraging others and giving support
5) Group Belonging: involves a sense of belonging and group identity	5) Social Objects: Topics about social objects
6) Self-Escape: Topics related to escapism or self-preservation	6) Seeking Support: Topics that deal with seeking support and help from others
7) Improving the current situation: Topics about improving the current situation or seeking change	7) Emotional states: Topics related to emotions and affective states
8) Personality Type Testing: Topics related to testing and assessing an individual's personality type	8) Work life: Covers topics related to work and life

we computed the cosine similarity between all topic words within two topics, subsequently deriving the average similarity score. For the topic $\{A_1, \dots, A_n\}$, each topic has m keywords, and for the i -th topic, the keywords are $\{w_1^i, \dots, w_n^i\}$, the BERT model is f , and the similarity between the i -th and j -th topics is:

$$R_{ij} = \frac{1}{m^2} \sum_k \sum_t \frac{f(w_k^i) f(w_t^j)}{\|f(w_k^i)\| \|f(w_t^j)\|} \quad (5)$$

Figure 2 reveals a notable distinction in topic similarity between the two virtual communities. Specifically, the group 1 virtual community exhibits a topic similarity score of 0.256, indicating a comparatively higher degree of topic concentration within its posting content. Conversely, the group 2 virtual community records a topic similarity score of 0.223, signifying a relatively greater dispersion of topics within its content. Topic similarity values range from 0 to 1, with

higher values indicating more similar topics. In Group 1, high topic similarity suggests common subject matter and focused discussions. Group 2, on the other hand, shows low topic similarity, indicating diverse discussions spanning various interests. This insight helps us understand community characteristics and topic distribution, guiding further research on member behavior and interactions.

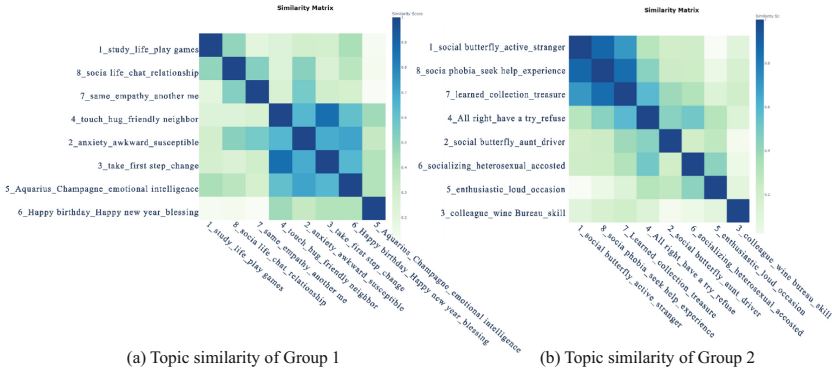


Fig. 2. Topic similarity Analysis

4.2 Output of Semantic Network Analysis

Based on our text topic analysis, we identified and defined 8 distinct clusters. Figure 3 visually displays high-frequency words in each cluster. To measure internal network cohesion, we used the modularity metric. Introduced by Newman and colleagues, modularity assesses network division quality. It's computed based on connection strengths among nodes and community partition consistency. Higher connection strength between nodes in the same community yields a greater modularity value.

Modularity values are typically bounded between -1 and 1, with values approaching 1 indicating stronger internal connections within communities and more effective community delineations. We first divide a social network in to k communities and defines a $k \times k$ matrix e , where e_{ij} represents the proportion of edges connecting community i and community j relative to the total number of edges. let $a_i = \sum_j e_{ij}$

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - |||e^2| \tag{6}$$

Utilizing the aforementioned definition, we calculated the modularity scores for both group 1 and group 2, resulting in values of 0.565 and 0.541, respectively. With the above definition, we can obtain the modularity of group 1 and group 2, which is 0.565 and 0.541 respectively.

The semantic network analysis of Fig. 3 reveals complex relationships. In Group 1’s network, prominent themes include “interpersonal” and “emotion.” In the interpersonal network, scenarios like schools and companies are linked to emotional attributes like tension and dislike. While both Group 1 and Group 2 focus on “emotion” in their networks, there are notable differences in associated words. Group 2 includes positive terms like “happy,” “courage,” and “possess,” whereas Group 1 uses more negative words like “introverted,” “difficult,” and “inferiority.”

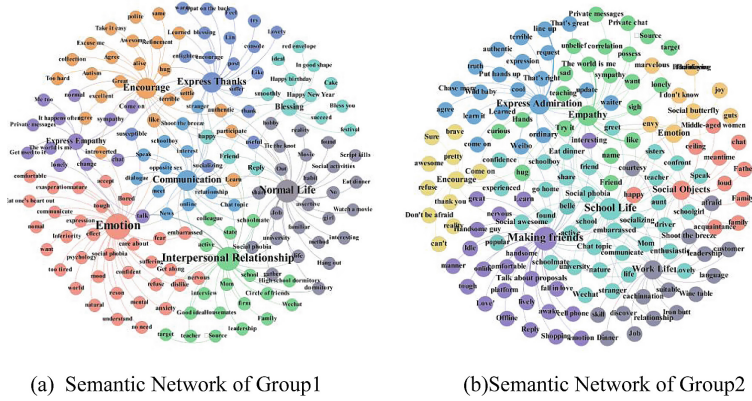


Fig. 3. Semantic Network Visualization of Virtual Community Modules

Discourse analysis highlights family relationships’ impact on teenage social anxiety, reflected in high-frequency words like “home,” “parents,” and “child.” Frequent use of words such as “communication,” “change,” “try,” “method,” “way,” “advice,” “want to,” and “chat” indicates active support-seeking and offering by members, including both information and emotions. In Group 2, interactions mainly involve strangers, evident in discussions related to unfamiliar individuals like “aunt,” “driver,” and unknown “middle-aged women.” Group 2 members tend to engage with a broader social circle, contrasting with Group 1’s behavior.

4.3 Outcomes of Sentiment Contagion

Sentiment contagion explores emotions’ diffusion and impact in social networks, enhancing our grasp of online emotional dynamics. It describes how individuals or groups’ moods are influenced by others, especially in social media, leading to emotional spread, creating emotional homophily [19]. Sentiment contagion is complex, influenced by factors like network evolution [21], individual susceptibility [3, 14], user engagement levels [12], and the sway of opinion leaders [3]. Prior research [8] indicates that emotions expressed in main posts affect members’ responses. This study explores how emotions like happiness and sadness in main posts trigger corresponding responses, impacting values, cognition, and

behavior. It confirms the impact of main post sentiment on replies. Additionally, this study reveals that different emotional posts receive varying attention in Douban Forum. Posts with more replies exert a stronger influence due to platform mechanics. Thus, the study examines the relationship between main post emotions, reply count, and reply emotions, as depicted in Fig. 4.

The experimental findings reveal that in Group 1, no significant link is found between main post emotions and reply tones, implying rational discussions prevail. In contrast, Group 2’s results highlight the impact of emotional value on response levels ($R^2 = 0.03$, $P < 0.001$). Positive emotions in Group 2 main posts elicit more responses, indicating a focus on emotional exchanges.

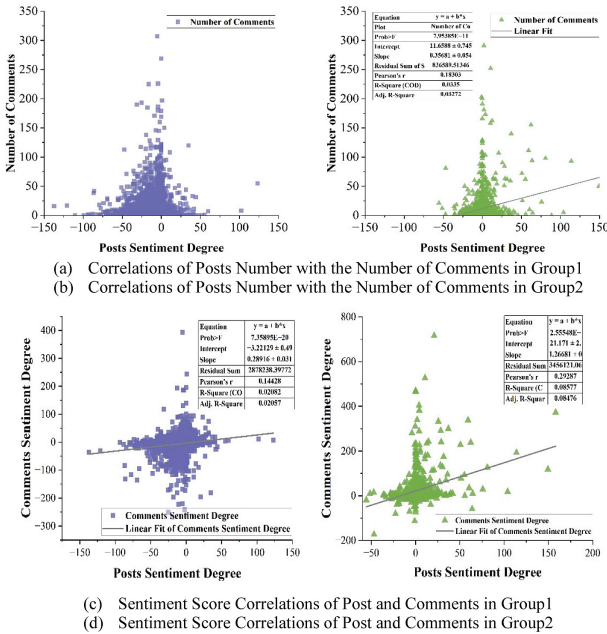


Fig. 4. Post and Comments index Relationship Diagram

Both groups display a meaningful positive correlation between reply and main post emotions. Negative main post sentiments lead to negative replies, showcasing sentiment contagion within social networks. Sentiment contagion occurred in both groups. In Group 2, positive main posts create a positive emotional environment, while Group 1’s negative posts lead to a more pessimistic atmosphere.

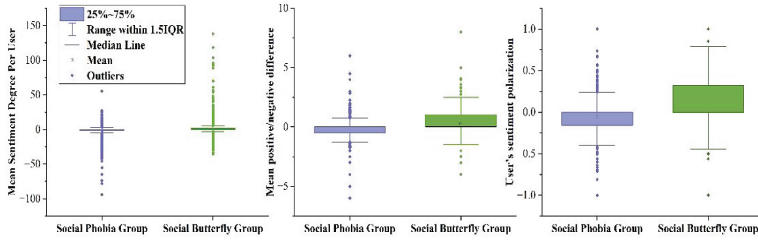
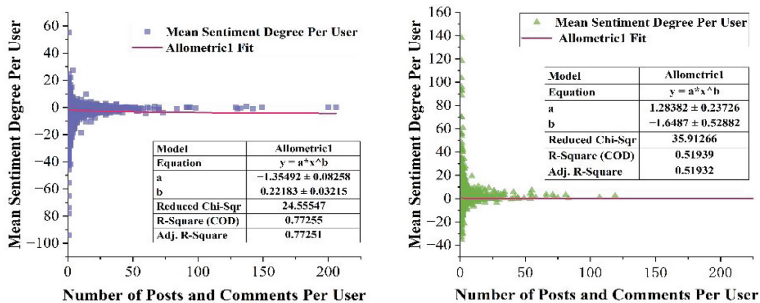


Fig. 5. Figure Distribution of three emotional indicators in Group 1 and Group 2

To enhance understanding of emotional differences, we introduce three sentiment indicators: average user sentiment, positive/negative comment difference, and sentiment polarization, as shown in Fig. 5.

The analysis of Fig. 5 reveals Group 2's dominance over Group 1 in all three emotional indicators. This suggests that while rational behavior is common to both communities, Group 2 members are more influenced externally, evident in their shifts from negativity to positivity. This susceptibility highlights Group 2's stronger sentiment contagion, possibly emphasizing community nature's role over structure in emotional changes. Group 2's superior results across all three indicators emphasize the prevalence of positive emotions, attributed to interaction patterns, emotional support, and reinforcement mechanisms. These findings stress that a community's nature, culture, and values significantly impact emotional transformations in social networks.

This study reinforces the importance of community nature in shaping emotions within online communities. To understand the relationship between user sentiment and activity, we analyzed the average user sentiment against the total comments posted. Figure 6 depicts the results. A clear link between user activity and their emotional state emerges. When users post fewer comments, their average sentiment tends to be lower. However, as they become more active, their aver-



(a) Correlation between sentiment scores and number of comments in Group 1

(b) Correlation between sentiment scores and number of comments in Group 2

Fig. 6. Correlation analysis between sentiment scores and number of comments

age sentiment rises, indicating a positive correlation. This suggests that greater activity correlates with increased social interaction, emotional expression, and improved emotional experiences for users.

Group 1 members exhibit predominantly negative emotions, fostering the spread of negativity within their community. Group 2 is more varied, with two polarized personality types. Negative sentiment in Group 2 is somewhat correlated with participant activity. Despite the negativity, most active Group 2 members maintain positive sentiment. To explore the link between positive/negative comment sentiment difference and user activity, we analyze emotional shifts with total posts and replies as variables in Fig. 7 displays the findings.

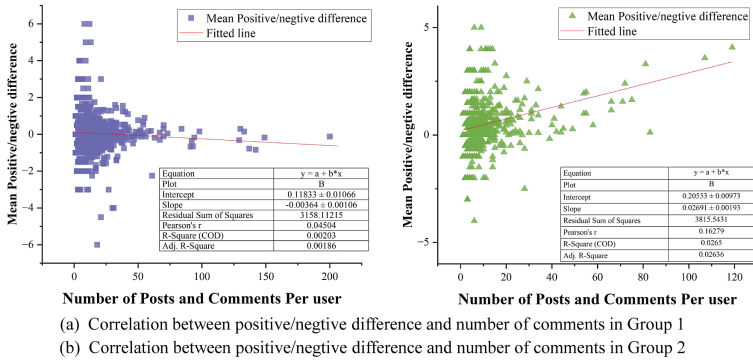


Fig. 7. Correlation Correlation analysis

Results revealed differing patterns in the two groups. In Group 1, increased engagement correlated with less efficient transitions from negative to positive attitudes. In contrast, Group 2 showed that higher activity linked to more efficient shifts from negative to positive attitudes. Group 1 active users tended to shift from positive to negative attitudes. This highlights that the community's prevailing emotional atmosphere significantly influences members' emotions. In Group 2, active users tend to shift from negative to positive attitudes due to the positive community ambiance. Conversely, in Group 1, active users are more influenced by the negative community atmosphere, leading to transitions from positive to negative attitudes.

This emphasizes the importance of the community's emotional environment in shaping users' emotional shifts. We examined the relationship between user sentiment polarity and activity, using user sentiment polarity as the dependent variable and members' total posts and replies as the independent variable. See Fig. 8 for results.

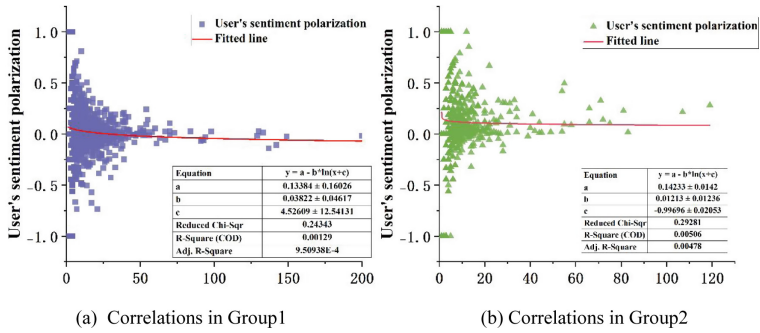


Fig. 8. Correlation Correlation between sentiment polarity and the total number of posts

Experimental results show that a sentiment polarity of 0 occurs when members balance positive and negative posts or share neutral content. As user activity increases, community members' sentiment polarity becomes more distinct. This emotional orientation results from a combination of community attributes and the prevailing atmosphere. In Group 1, where content leans negative, heightened activity corresponds to a more negative community sentiment. In Group 2, with initially positive sentiment polarity, increased activity tends to shift it towards neutrality. Further analysis of Group 2 reveals that beyond a certain threshold ($n > 50$ comments), sentiment polarity starts leaning positive with increased participation. This highlights the relationship between user activity and sentiment polarity. Community attributes and overall atmosphere play vital roles in shaping sentiment polarity among different groups.

5 Discussion and Conclusion

5.1 Discussion

This research leverages advanced technologies, including BERTopic, NLP, and social network analysis, to explore emotional communication and transformation in virtual communities. BERTopic aids in uncovering sentiment orientations and latent themes in social network data. By analyzing social media text, we reveal how user emotions spread and information flows in social networks.

We employ sentiment analysis algorithms to automatically detect and quantify emotional polarity in user posts and comments. Integrating social network methods with time series analysis helps us examine sentiment propagation's speed and impact. Text clustering techniques extract keywords and emotional discourse themes. This research delves into collective intelligence phenomena, providing a holistic understanding of emotional interactions and information dissemination. Challenges include handling diverse expressions and complex emotional nuances in social network texts.

The findings highlight community attributes, ambiance, activity, and participation as collective influencers on emotion diffusion in social networks. The study enhances our comprehension of emotional transmission mechanisms, offering insights for fostering positive community atmospheres.

However, limitations exist in this methodology, including potential for larger sample sizes, diverse data sources, and integrating qualitative methods. Investigating platform design and algorithmic effects on emotional communication could lead to more effective social network management strategies.

5.2 Conclusion

This study represents a comprehensive investigation into the intricate dynamics of emotion propagation and transformation within social networks. It highlights the value of combining BERTopic and NLP for sentiment analysis in social networks. Main posts significantly influence reply sentiments, while community atmosphere affects sentiment spread. User activity and emotional disposition are complexly linked; high activity often corresponds to negative emotions. BERTopic, NLP, and social network tech are vital for understanding emotional interactions and collective intelligence.

The implications of these findings extend beyond academic research. They hold practical relevance in the domains of emotion management and public opinion analysis within social network platforms, offering innovative methodologies to delve into the mechanisms governing the dissemination of emotions and the dynamics of collective intelligence behaviors within the realm of social networks.

References

1. Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., Quattrociocchi, W.: Echo chambers: emotional contagion and group polarization on facebook. *Sci. Rep.* **6**(1), 37825 (2016)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
3. Doherty, R.W.: The emotional contagion scale: a measure of individual differences. *J. Nonverbal Behav.* **21**, 131–154 (1997)
4. Liu, G., Wang, T., H.T.K.W.Y.: Enhanced contextual neural topic model for short text. *Computer Engineering and Applications*, pp. 1–14 (2023)
5. Gómez-Ortiz, O., Roldán, R., Ortega-Ruiz, R., García-López, L.J.: Social anxiety and psychosocial adjustment in adolescents: relation with peer victimization, self-esteem and emotion regulation. *Child Indic. Res.* **11**, 1719–1736 (2018)
6. Gómez-Ortiz, O., Romera, E.M., Jiménez-Castillejo, R., Ortega-Ruiz, R., García-López, L.J.: Parenting practices and adolescent social anxiety: a direct or indirect relationship? *Int. J. Clin. Health Psychol.* **19**(2), 124–133 (2019)
7. Grootendorst, M.: Bertopic: neural topic modeling with a class-based tf-idf procedure. arXiv preprint [arXiv:2203.05794](https://arxiv.org/abs/2203.05794) (2022)
8. Guangyang: An empirical analysis of social media network emotional contagion and clue influence mechanism. *J. Shenzhen Univ. Humanities Soc. Sci. Edition* **37**(6), 12 (2020)

9. Hofmann, S.G., Anu Asnaani, M., Hinton, D.E.: Cultural aspects in social anxiety and social anxiety disorder. *Depress. Anxiety* **27**(12), 1117–1127 (2010)
10. Islam, M.R., Kabir, M.A., Ahmed, A., Kamal, A.R.M., Wang, H., Ulhaq, A.: Depression detection from social network data using machine learning techniques. *Health Inform. Sci. Syst.* **6**, 1–12 (2018)
11. Jiang, S., Ngien, A.: The effects of instagram use, social comparison, and self-esteem on social anxiety: a survey study in Singapore. *Soc. Media+ Soci.* **6**(2), 2056305120912488 (2020)
12. Kramer, A.D., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U.S.A.* **111**(24), 8788 (2014)
13. Melka, S.E., Lancaster, S.L., Adams, L.J., Howarth, E.A., Rodriguez, B.F.: Social anxiety across ethnicity: a confirmatory factor analysis of the fine and sad. *J. Anxiety Disord.* **24**(7), 680–685 (2010)
14. Meng, L.M., Duan, S., Zhao, Y., Lü, K., Chen, S.: The impact of online celebrity in livestreaming e-commerce on purchase intention from the perspective of emotional contagion. *J. Retail. Consum. Serv.* **63**, 102733 (2021)
15. Nor, N.M., Rahman, N.A., Yaakub, M.R., Zukarnain, Z.A.: Sentiment analysis on depression detection: a review. In: *Science and Information Conference*. pp. 718–726. Springer (2022)
16. Pengguan, Wang, F.: Research on the optimal topic number determination method of lda theme model in scientific and technological information analysis. *Modern Library and Information Technology* (9), 9 (2016)
17. Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R.: How to analyze political attention with minimal assumptions and costs. *Am. J. Political Sci.* **54**(1), 209–228 (2010)
18. Ruz, G.A., Henríquez, P.A., Mascareño, A.: Sentiment analysis of Twitter data during critical events through bayesian networks classifiers. *Futur. Gener. Comput. Syst.* **106**, 92–104 (2020)
19. Schoenewolf, G.: *Emotional contagion: Behavioral induction in individuals and groups*. Modern Psychoanalysis (1990)
20. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of depression-related posts in reddit social media forum. *Ieee Access* **7**, 44883–44893 (2019)
21. Xiong, X., Li, Y., Qiao, S., Han, N., Wu, Y., Peng, J., Li, B.: An emotional contagion model for heterogeneous social media with multiple behaviors. *Phys. A* **490**, 185–202 (2018)
22. Zhou, Y., Xu, R., Gui, L.: A sequence level latent topic modeling method for sentiment analysis via cnn based diversified restrict boltzmann machine. In: *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 356–361. IEEE (2016)